UNIVERSIDADE DE LISBOA
FACULDADE DE CIÊNCIAS
DEPARTAMENTO BIOLOGIA ANIMAL

Building a Portal for Scientific Collections at the University of Lisbon

Pedro Miguel Oliveira Ladino

Mestrado em Bioinformática e Biologia Computacional

Dissertação orientada por:
Professora Doutora Cátia Luísa Santana Calisto Pesquita
Doutora Maria Cristina Duarte

2020

Acknowledgements

Firstly, I must thank professor and supervisor of this project Cátia Pesquita and professor and cosupervisor Maria Cristina Duarte for their invaluable input of knowledge, patience, confidence and motivation given to me throughout the entire duration of this work, even when times looked dark. Professor Dulce Domingos must also be thanked for her relentless follow-up on this project and all the help she provided me inside the Reitoria da Universidade de Lisboa and with the communication with MUHNAC. Gratitude to the MUHNAC personal, in particular to Dr. Marta Lourenço, Dr. Alexandra Cartaxana, Dr. Inês Pinto and Dr. Judite Alves for providing me with the help, documentation and suggestions used in this work. I am obliged to the Fundação para a Ciência e Tecnologia for their research Grant, made available to me via funding from LASIGE Strategic Project (UIDB/00408/2020) and by the SMILAX project (PTDC/EEI-ESS/4633/2014)

Finally, I extend my deepest gratitude towards my family who stood with me during the best and worst of times, and my fellow scholarship friends Diogo Godinho, Diana Paiva, Francisco Caeiro, Mafalda Luz and Nuno Ricardo for the countless coffee breaks and philosophical discussions that allowed me to take a mental reset when things were not going my way.

I

Resumo Alargado

Nos últimos anos, tem se verificado um desenvolvimento exponencial nas técnicas utilizadas para estudar as coleções científicas e os dados que a elas estão, geralmente, associados. Tais avanços resultam na produção de novos dados e novas propriedades, associadas aos mesmos, que têm de ser guardados, anotados e mantidos juntamente com os dados previamente existentes, resultando numa constante necessidade de adaptar e melhorar a capacidade de gerir, manter e valorizar esta enorme quantidade de informação.

As coleções científicas, integrando objetos recolhidos para estudo e toda a informação que lhes está associada, englobam diversos domínios como a Botânica, a Zoologia, a Geologia, a Microbiologia, entre muitos outros. Os dados das coleções científicas são agrupados, de acordo com os seus domínios e informação comum, e são preservados, catalogados e manuseados com o objetivo de permitir o seu posterior uso para efeitos de pesquisa ou de divulgação.

Estas coleções, na maior parte das vezes, estão sobre a responsabilidade de museus e dos seus respetivos curadores e, por isso, é importante que exista um suporte tecnológico que permita, de uma maneira rápida e eficiente, o manuseamento, manutenção e partilha destes dados. Várias iniciativas internacionais surgiram nas últimas décadas com o objetivo de contribuir para a partilha de dados e catalogação dos mesmos, como o Global Biodiversity Information Facility (GBIF), que fornece uma arquitetura onde diferentes museus e curadores podem submeter os seus dados, ficando estes acessíveis globalmente, ou o Integrated Digitized Biocollections (IDigBio), que se foca na ajuda aos processos de digitalização dos dados.

O principal objetivo deste projeto era disponibilizar ao Museu Nacional de História Natural e da Ciência (MUHNAC) da Universidade de Lisboa uma plataforma web que os diferentes curadores pudessem utilizar para o tratamento e manutenção dos dados das coleções de que são responsáveis. Estes dados têm incidências sobre múltiplos domínios das áreas das ciências, resultando numa diversidade de coleções com diferentes propriedades e características que têm de ser exploradas e implementadas. A padronização dos dados de maneira a que estes sigam descrições e propriedades utilizadas internacionalmente também foi um dos grandes objetivos, com a ontologia Darwin Core (DwC) a servir de referência. O software open-source CollectiveAccess (CA), desenhado para a manutenção de coleções, foi escolhido para servir de base à plataforma web desenvolvida. Utilizando as funcionalidades base do CA, em conjunto com ficheiros de código produzidos durante o projeto, foi possível criar um sistema que atende a uma grande parte dos desafios presentes na manutenção de dados científicos, como a implementação de um sistema que gere dados de origem taxonómica, dados geográficos e de geolocalização, individualização das interfaces de inserção de dados, customização dos formulários de pesquisa atendendo às preferências dos curadores e acessos discriminatórios aos diferentes dados.
Foi, também, desenvolvida uma metodologia que visa generalizar todo o processo desde a extração dos dados da sua origem, tratamento de correções ortográficas e produção de um modelo de dados que siga ao máximo, e sempre que possível, os termos descritos pela ontologia DwC, até à sua importação para o CA, passando pelo desenho e criação de ficheiros de mapeamento, a sua importação para o sistema e desenho de interfaces que mostrem as propriedades de cada conjunto diferente de dados.

Para avaliar o comportamento da plataforma e as funcionalidades implementadas, foi desenhado um conjunto de testes de usabilidade que incidiam sobre as operações permitidas a dois tipos de utilizadores expectáveis do sistema; a função de curador e a função de utilizador. Ambos os testes incluíam a

realização das mesmas operações de pesquisas e dos diferentes tipos de pesquisa no sistema. O teste dos curadores incluía, ainda, uma secção com foco na inserção de um novo registo escolhido criteriosamente, de maneira a que as funcionalidades implementadas estivessem presentes e pudessem ser testadas. A avaliação dos testes foi efetuada recorrendo à análise de um formulário de System

Usability Scale (SUS), tendo sido obtido um resultado médio de 80 para os testes com utilizadores públicos e 77,25 para os testes dos curadores, classificando o projeto com uma nota A e B, respetivamente, numa escala de usabilidade. Foram, também, pedidas sugestões e comentários de modo a avaliar o funcionamento do sistema e possíveis futuras funcionalidades ou mudanças a serem implementadas.

Por fim, cabe realçar que com este projeto foi possível implementar uma primeira abordagem a um sistema para gestão, manutenção e pesquisa de dados associados para todas as coleções pertencentes ao MUHNAC num lugar único e em que standards internacionais, em termos de metadados (DwC), foram seguidos e com possibilidade de exportação dos dados em formatos relevantes para os diferentes projetos internacionais (e.g., GBIF). É, também, esperado que este projeto facilite consideravelmente a manutenção e gestão integrada das coleções do MUHNAC e, simultaneamente, forneça um suporte valioso para o futuro uso dos dados nas atividades de investigação científica.

Resumo

As coleções científicas, reunindo uma enorme quantidade e diversidade de objetos e os dados que lhes estão associados, constituem um valioso património histórico, científico e cultural. Estas coleções estão, geralmente, sob a responsabilidade dos museus e dos seus respetivos curadores, sendo importante que exista uma plataforma sobre a qual os responsáveis das mesmas possam efetuar operações de gestão e de manutenção das mesmas.

Atendendo à diversidade das coleções, estes dados, pertencentes a diferentes domínios científicos e com propriedades distintas, colocam problemas de integração, disponibilização e manutenção, problemas estes cada vez mais pertinentes numa realidade que vive de dados e da análise e partilha dos mesmos.

Este projeto, centrado neste desafio, pretendeu desenvolver, para o Museu Nacional de História Natural e da Ciência da Universidade de Lisboa, uma plataforma que agregasse as variadíssimas coleções desta instituição, tirando partido de uma plataforma *open-source* base chamada CollectiveAccess. No decorrer do mesmo, foi desenvolvida uma metodologia generalizada para qualquer coleção que cobre os processos desde a aquisição dos dados, o seu processamento e correção até à sua importação e disponibilização dentro da plataforma. Foram, também, desenvolvidas e implementadas funcionalidades específicas que visaram resolver determinadas características particulares dos diferentes conjuntos de dados como é o caso da implementação de um sistema hierárquico para dados relacionados com taxonomia, sistema de introdução de dados geográficos utilizando uma API externa e desenvolvimento das funcionalidades de pesquisa de modo a satisfazerem as necessidades de cada conjunto de dados.

Estas funcionalidades e o desempenho do sistema foram avaliados através de dois questionários de usabilidade (System Usability Scale), através de dois Google Form diferentes. Estes questionários foram direcionados para dois tipos principais de utilizadores do sistema: curadores e público, em geral. Para além disto, foram pedidos comentários e sugestões de melhorias ou acrescento de funcionalidades. Os resultados dos questionários foram satisfatórios obtendo-se uma classificação de A e B, por parte dos testes do público e dos curadores respetivamente, na escala de usabilidade. A análise dos comentários e sugestões também permitiu obter uma ideia sobre possíveis melhoramentos e novas funcionalidades a implementar.

Palavras-chave: Portal-Web, Coleções Científicas. Museus, Darwin-Core,

Abstract

With scientific collections bringing together a huge number and diversity of objects and the data associated with them, they constitute a valuable historical, scientific and cultural heritage. These collections are generally under the responsibility of museums and their respective curators, and it is

important that there is a platform on which those responsible for them can carry out management and maintenance operations.

Given the diversity of the collections, these data, belonging to different scientific domains and with different properties, pose problems of integration, availability and maintenance, problems that are increasingly relevant in a data-centric world that relies on the analysis and sharing of the data.

This project, focused on this challenge, aimed to develop, for the Museu Nacional de História Natural e da Ciência da Universidade de Lisboa, a platform that aggregates the very diverse collections of this institution, taking advantage of an open-source base platform called CollectiveAccess. In the course of the same, a generalized methodology was developed for any collection, covering the processes from the acquisition of the data, its processing and correction to its import and availability within the platform. Specific features were also developed and implemented that aimed at solving certain particular characteristics of different data sets, such as the implementation of a hierarchical system for taxonomyrelated data, geographic data entry system using an external API and development of the base search features, meeting the requirements for each collection.

These functionalities and the overall performance of the system were evaluated through two usability questionnaires (System Usability Scale), via two different Google Forms. These questionnaires were aimed at two main types of users of the system: curators and the general public. In addition, comments and suggestions for improvements or addition of features were requested. The results of the questionnaires were satisfactory, obtaining a classification of A and B, by the tests of the public and the curators, respectively, on the usability scale. The analysis of comments and suggestions also provided an idea of possible improvements and new features to be implemented.

Keywords: Web-Portal, Scientific Collections, Museums, Darwin-Core

Contents

List of Figures

List of Tables

1 Introduction

1.1 Motivation

As science and scientific discoveries grows larger by the day, there is a never-ending demand to store the information that is newly discovered alongside the information we already had.
This huge influx of new data, alongside its metadata, present us with new challenges when it comes to integrating it alongside the data already stored and catalogued, performing proper annotations of the data to increase search efficiency, legal and ethical hurdles on data sharing, data storage and secure backups, amongst other [1].

A scientific collection is essentially a group of related scientific objects, sharing a number of common features, that are intended to be preserved, managed and catalogued for the purpose of allowing future studies [2]. They are also cultural objects, documenting history and heritage assets of the naturalists and other explorers that travelled around the world [3].
Scientific collections cover the usual Botanical and Zoological subjects, but they go much further than just these two big domains. Geology, Paleontology, DNA Banks, microorganisms, rock cave pictures, old transcripts from ancient times are just a few examples of record domains also included in scientific collections [2,3].

These collections are maintained by a variety of institutions ranging from natural history museums, botanical gardens, universities and other research institutions. For instance, the Natural History Museum of London, holds one of the biggest collections worldwide, with over 80 million items within five main collection areas: Botany, Entomology, Mineralogy, Paleontology and Zoology [4]. Natural History museums should be considered as critical infrastructures for scientific inquiry and public understanding, playing a fundamental role in our societies [5].

A prominent question about scientific collections is the reason to why people should care about them. In reality, and often unrealized and discretely, scientific collections provide valuable knowledge. Each specimen can provide many kinds of data (e.g., information on locality and other biotic and abiotic collection parameters, be used as source of DNA or other molecular materials, just to mention a few examples). This wealth of metadata turns scientific collections into powerful research tools, enabling scientists to test hypotheses and carry out varied studies [5].
Also, they have been widely used to the management and governments decision-making. Accordingly, to a report of the Interagency Working Group on Scientific Collections [2] several topics were reported to be majorly impacted by scientific collections: e.g., in economy and trade (foreign and domestic trades are supported by research); environmental quality (in modelling future environmental changes so they can be better managed); controlling and preventing invasive species (food and parasites control at borders); public health and safety (diseases study and forecast of new epidemics, drug discovery and drug testing using collection data); and many other unanticipated uses, possible by the continuous development of new analytical techniques, and allowing researchers to ask new and more detailed questions using the same collections. Moreover, and mostly, scientific collections contain unique objects that cannot be collected again easily or at all, making them priceless.

Scientific Collection databases

In biological sciences fields, for instance, huge efforts have been made to store scientific collections data in databases. A report from the OECD Megascience Forum Working Group on Biological Informatics [6] stresses the importance of making biodiversity data and information accessible worldwide. Many initiatives have evolved in recent years to help with the constant growing number of biological data as well as the demands and standards we have for data cataloging and sharing of that data. Some examples are:

-The *Global Biodiversity Information Facility* (GBIF), launched in 2001, is a very well-known project that tries to tackle the accessibility in the sharing of scientific biodiversity data via Internet. GBIF basically provides and architecture in which different collection holders can submit their data and make it accessible and searchable through a single web-portal, facilitating the free and open access to scientific data [4].

-The *Integrated Digitized Biocollections* (IDigBio), launched in 2011 it's one of those initiatives that is playing a role in helping with the digitalization of millions of scientific data and making them available for community (research community, governments, agencies, students, educators and general public) [4].

- The *Distributed System of Scientific Collections* (DiSSCo), launched in 2018, is a Research Infrastructure for natural science collections that aims to unify natural science collections in Europe by allowing Europe's researchers and technology professional to share and reuse the data linked to collections across disciplines and borders. Once again providing a unique and centralized infrastructure where the collection's data is stored [4].

This being said, it is important for institutions that are responsible for holding these scientific collections to be able to maintain them, cataloguing them according to international standards and, moreover, to fully develop ways they can be easily accessed by the whole scientific community and general public.

This context brings us to the focus of this dissertation, which is to empower the National Museum of Natural History and Science from the University of Lisbon (MUHNAC) with a sustainable and publicly accessible web repository, so that the invaluable scientific collections of this institution can continue to play a major role in promoting and improving the scientific knowledge, and, simultaneously supporting management tasks.

Natural History Collections at the MUHNAC

The MUHNAC has a huge variety and diversity types of collections, including botanical, zoological, geological, paleontology, only to mention a few, and the number of items belonging to those is estimated to far surpass 1 million records (Marta Lourenço, Director of MUHNAC, pers. com.).
These collections represent a major source of scientific information that must be preserved and thoroughly documented in order to support research studies and national or foreign collaborations. Having such big datasets, that embrace not only the natural history of the Portuguese territories, but also that of many worldwide regions (namely the Portuguese-Speaking countries), presents a great opportunity for analysis and exposure of this invaluable scientific heritage, not only inside Portugal but also internationally.

The efforts made in last years have already produced databases for some of the Museum collections. However, a main problem is that there is no uniformization and standards when it comes to where all the MUHNAC data is stored, in which system and format and how the data is described via their metadata. Some data lives inside excel spreadsheets, other in third-party software or even external isolated databases using different database languages (e.g., Specify, Access, FileMaker, MySQL). This project is an attempt to tackle this siloed, sparce and standards-uncompliant data and produce an integrated standards-compliant data model, with the information being stored at a single location with secure backups and made available through a web browser.

1.2 Goals

This work aimed to:

- Understand the requirements for an integrated management, valorization and use of the MUHNAC scientific collections.
- Analyze existing state of the art collection platforms to match the identified requirements.
- Producing a metadata model that is aligned with international vocabularies and standards.
- Adapt and extend a collection platform to support access and management of specimen collections at MUHNAC through a web portal.

1.3 Contributions

This work has made the following contributions:

- Provide the MUHNAC with a single central platform design to host all their data regardless of its domain.
- Make the data about the MUNHAC collections FAIR (findable, accessible, interoperable and reusable).
- Produce a metadata model aligned with the Darwin-Core ontology.

1.4 Document Structure

This document is structured as follows:

- Chapter 2 - State of the Art, provides an overall view of different collection management platforms and portals and outlines some of the most important features and the data standards and controlled vocabularies used in this project.

- Chapter 3 - Methodology, suffices a more detailed explanation about the different data domains, requirements and use cases, gap analysis, metadata and architecture of this project and the software used in it.

- Chapter 4 - Implementation, focus on the some of the main features or processes implemented.

- Chapter 5- Evaluation, analyses and discusses the results from a public usability test.

- Chapter 6- Conclusion and Discussion, presents the main conclusions and discussions and provides suggestions for future work.

2 State of the art

This chapter will focus on the state of the art when it comes to collection managements platforms and some of the most known international portals browsing scientific collections and its data. It will also cover the main ontology used throughout this project.

2.1 Platforms

There are plenty of software platforms for managing scientific collections available through the internet that cover different collections needs. Some of them are paid software like “PastPerfect”, “ArtLogic”, or “Artwork Archives” others are open source and free projects like “Specify”, “CollectionSpace” or “CollectiveAccess”. Overall, every single one of the platforms offers a different perspective and focus on how collections should be managed and what are some of the key features necessary. Some platforms are heavily focused on a specific type of collections, like “ArtLogic” that was heavily designed to handle Art collections, or “Specify” that was built as a biological collection management software; other platforms adopted a more general approach whereby compromising some more type specific features allowed for a broader range of collections types like “CollectionSpace” and “CollectiveAccess”.

Besides the types of supported collections, as a management software everything from object entry and object acquisition to inventory control, location and movement control, cataloguing description, conservation management, loans and borrows are some of the required functionalities in order to be considered as a successful management software for implementation in a museum.

2.2 Portals

Looking at the context of this project of producing a web-based portal for the National Museum of Natural History and Science to manage scientific collections and make the information they contain fully accessible to the public, it makes sense to look at some of the most popular and used web-based interfaces with the same goals. This project consisted of mainly two big components, the curator’s interface, where records are inserted and managed, and a public interface for consultation.

When looking at some of the main natural history collection portals, only the public browsing interface is available for exploring, since the management interface is exclusive to authenticated users. Amongst the most popular portals, *Atlas of Living Australia* (Figure 2.1), an online repository for Australia biodiversity, the *Naturalis Biodiversity Center*, a natural history museum and a research center for Dutch biodiversity (Figure 2.2), and the *Muséum National D’Histoire Naturelle* from Paris (Figure 2.3) were highlighted as being some of the most important collection portals worldwide.

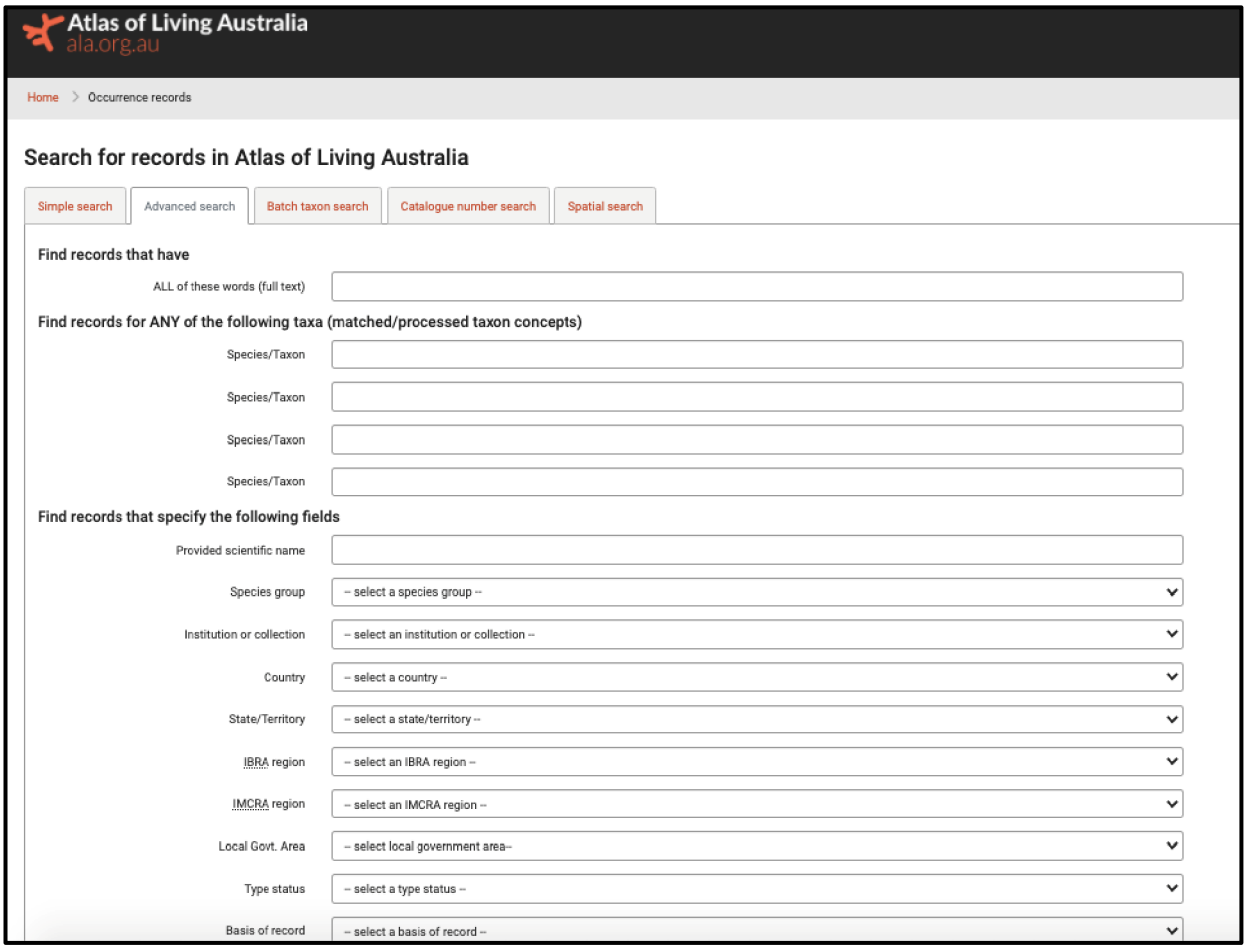

Figure 2.1- Atlas of Living Australia public Advanced Search Interfaces (partial image)

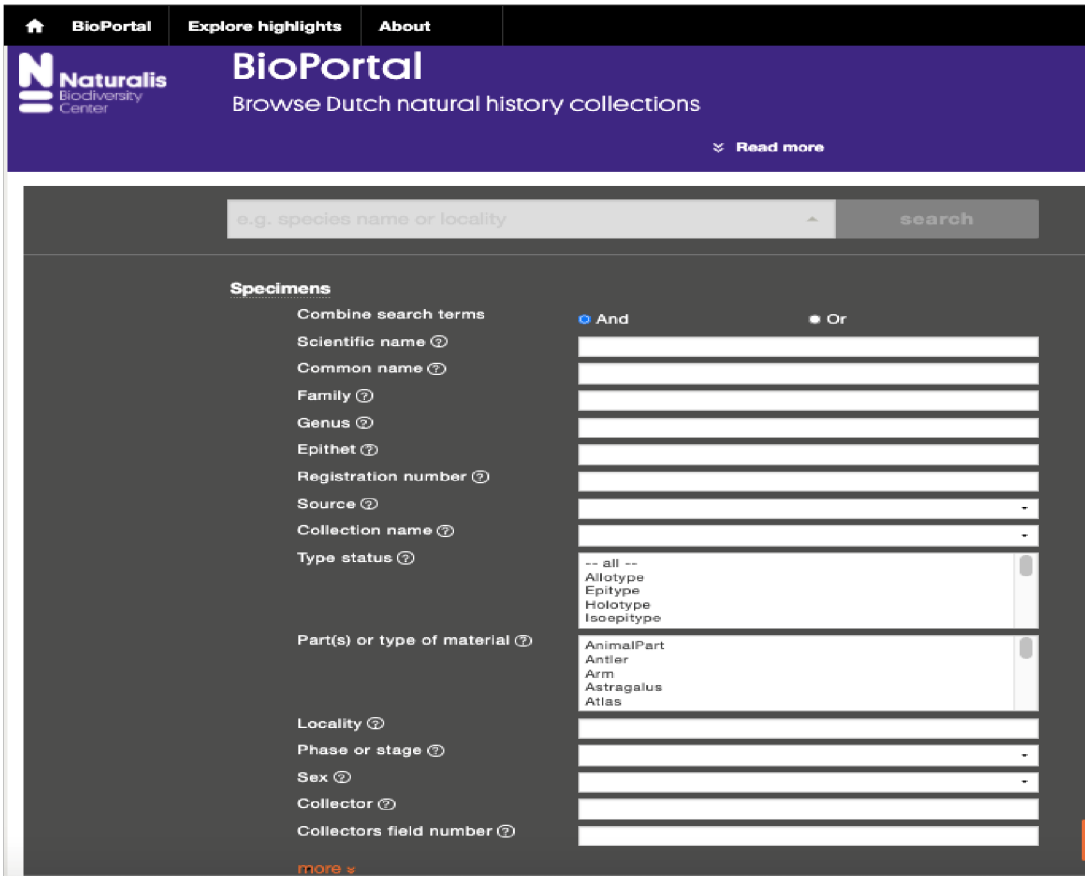

Figure 2.2- Naturalis public Advanced Search Interfaces (partial image)

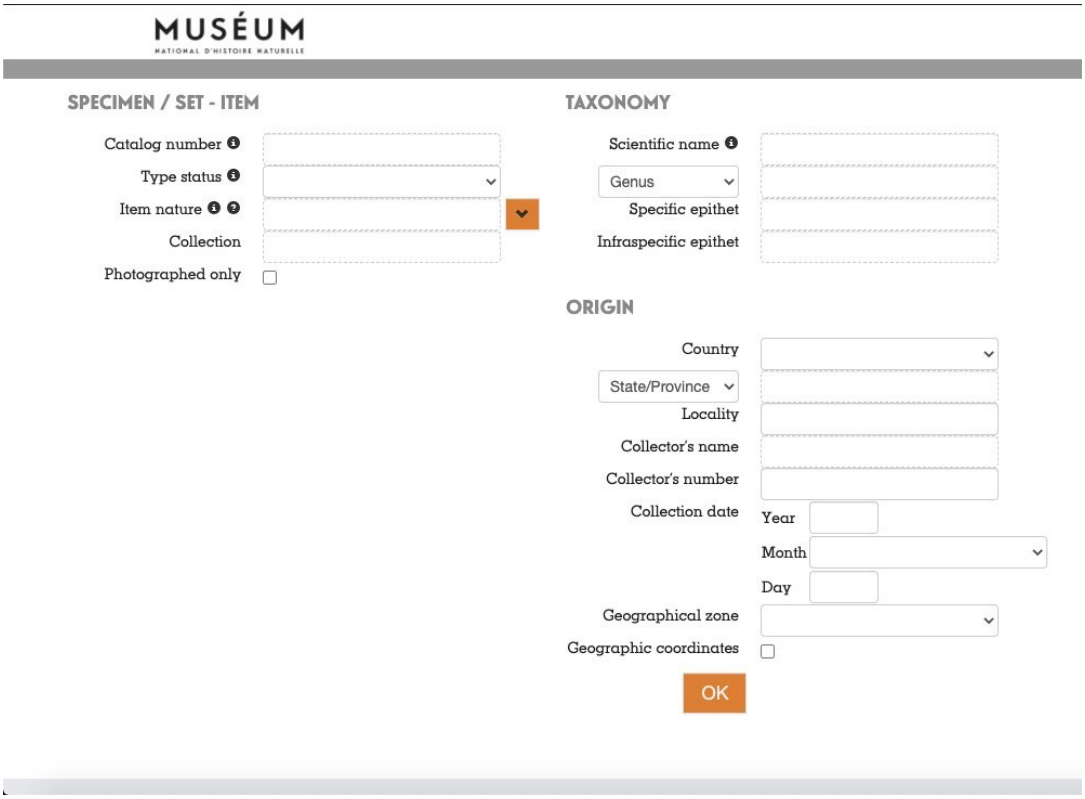

Figure 2.3- Muséum National D'Histoire Naturelle public Advanced Search Interface (partial image)

When looking at the public search interfaces for each one of them, some common features can be pointed:

- the multiple search options, either record specific options, collection specific options or even geographic options;
- the grouping of search options under different categories usually: Collection ID (…); specimen (including attributes as collector name and number, place and date of collection); and taxonomy (including several ranks, such as family, species and subspecies) amongst others.
- the possibility to combining multiple search fields into a single search for narrowing and getting more precise results.

Also, common features alongside these three different web portals are the clean lines of the interface, without any unnecessary information, the possibility to download the search results or the images associated with them, the visualization of site on a map (if a record has the required geographical information), and the metadata elements following the Darwin-Core Ontology.

This analysis together with the information transmitted by the potential users from the MUHNAC, during working meetings, helped to provide a first line drawing of the main features of how the MUHNAC search interface should look like and what are some of the most important search fields for public consulting.

2.3 Data standards and controlled vocabularies

It is a great challenge to make heterogeneous data sources interoperable and to unite them portals accessible for both the scientific community and the broader public. Specific data standards allow for an exchange and a standardized publication of collection object related data. Following a common standard schema, data from various institutions can be integrated, displayed and accessed via data portals in a sophisticated manner.

Access to Biological Collection Data

The data standard Access to Biological Collection Data (ABCD, Berendsohn 2007) is a well-known standard used for natural history collection and observation data. It is an evolving comprehensive standard for the access to and exchange of data about specimens and observations (a.k.a. primary biodiversity data) [9]. ABCD is currently in use with the GBIF (Global Biodiversity Information Facility) and BioCASe (Biological Collection Access Service for Europe) networks.

Darwin Core

Darwin Core (DwC) is an extension of Dublin Core for the fields of biodiversity informatics. Its intended to provide a stable standard for referencing and sharing information on biological diversity [9].

More specifically DwC is a set of standards that include a glossary of terms also usually referred as properties, elements, columns or attributes, intended to ease the share of information about biological diversity by providing reference definitions, examples and commentaries, making it easy to compare international samples. This standard describes biological samples mainly based on their taxonomical classification and occurrence in nature as documented by observations, specimens, samples and related information, and was the reference ontology to describe the data used in this project. In the latest DwC version there are 172 different terms to describe a biological sample [10].

3 Methodology

This project was developed in close collaboration with several stakeholders, listed below.
During the entire duration of this project approximately 12 meetings, involving at least this project's supervisors and a representative from the MUHNAC, took place and the requirements, progress and next steps were debated. This means there was a constant evolution in terms of requirements, and there was an iterative process of development, where refinement of requirements was born out of a joint analysis with the MUNHAC team and their formative evaluation of the developed approaches.

.

This chapter will focus on presenting the stakeholders, which is followed by a description of the main tasks: i) the datasets and their domains used throughout this project; ii) the functional requirements and use cases for the different operations and functionalities the software should be able to perform: iii) a gap analysis between different collection management systems; iv) database structure and the general methodology from getting a dataset to its implementation into the system and backing it up.

3.1 Stakeholders Involved

In this section are described some of the entities involved in the acquisition of the functional requirements, including some particular people and the central services of the University of Lisbon.

-DI SC – The Central Services Department of Informatics of the University of Lisbon is responsible for managing and maintaining all the software and systems in use within the University of Lisbon.

-MUNHAC – The National Natural History and Science Museum of the University of Lisbon is the main client of this project, and this work was developed having in mind their requirements.

-Dr. Alexandra Cartaxana, Dr. Judite Alves and Dr. Inês Pinto –are three collection curators working for MUHNAC and they were the responsibles for serving as a bridge between the work being developed and the MUHNAC, providing all the internal MUHNAC information necessary as well as some of the datasets and requirements.

-Dr. Maria Cristina Duarte - Was one of my supervisors during this project and she is also a curator for the MUHNAC, again bridging the gap with the Museum and providing me the data and functionalities used for testing the implementation of the requirements.

-Dr. Marta Lourenço – Is the director of the MUHNAC and has the overall final word on how the system needs to behave and the functionalities required.

Dr. Maria Dulce Domingos- Is one of the pro-deans of the University of Lisbon and provided invaluable and tireless amounts of support when it comes to using the infrastructure of the University of Lisbon for the implementation of this project as well as using her position to "pressure" the MUHNAC personal, so the necessary documents and data arrived on time.

3.2 Data and Domain Analysis

This section will cover the analysis of all the different collections owned by MUHNAC, as well as their domains and classification within the museum structure.

Data analysis

When we consider the dimensions and areas of science covered by the MUHNAC it should not come as a surprise the huge amount of data and data diversity they have. Analyzing a document provided by the Institution responsible that disclosures how their data is internally organized, we can see that there are 97 different collections from two different sources. There are the MUHNAC own collections and collections with integrated joint management by the MUHNAC and the former Instituto de Investigação Científica Tropical (IICT). Considering this number and that 26 different people are responsible for the majority of these collections (not including assistant curators, or collections that don't have a curator assigned to them), we can see the relevancy of some of the problems highlighted in the introduction such as data uniformization. In fact, each curator will have their own preferences in terms of metadata to describe their collections and data storage and is the solo responsible for where and how their information is stored.

For the purpose of this project, three datasets were used for testing, experimenting and building the system: The LISC Herbarium Collection, the Decapoda Collection, and the LISU Herbarium Collection, curated respectively by Dr. Maria Cristina Duarte, Dr. Alexandra Cartaxana and Dr. Judite Alves.

These three datasets (analyzed in more detail in the next chapter) provided over 200 different metadata elements with a combined amount of over 60000 records.

Domain analysis

Looking at the list of scientific fields/domains published by the Dutch Research Council (NWO) [11], we can consider the distribution of the MUHNAC (Figure 3.1) and MUHNAC+IICT (Figure 3.2) collections to be a part of the following main domains:

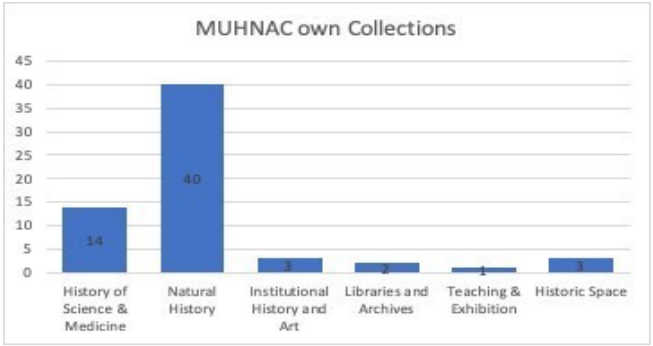

Figure 3.2- Number of MUHNAC collections grouped by their domains

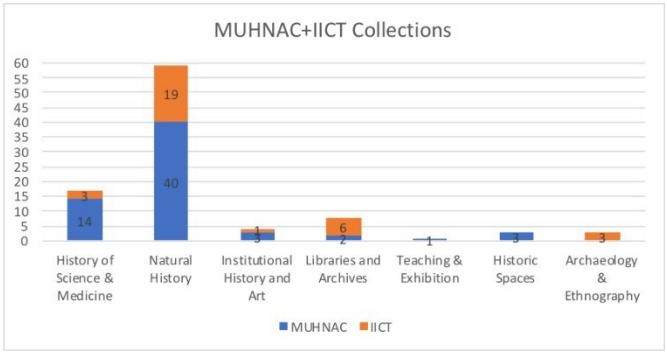

Figure 3.1- Number of MUHNAC and IICT collections grouped by their domains

In the overall distribution of domains, ~62% of the collections (59) can be considered as being part of the Natural History domain, ~18% of the collections (17) belong to the History of Science and Medicine domain while the remaining 20% collections (19) can be assigned to the Institutional History and Art, Libraries and Archives, Teaching and Exhibition, Historic Spaces and Archaeology & ethnography domains.

Digging deeper into the domain with the higher number of collections (Natural History), we can subdivide it into smaller artificial domains corresponding to the inner divisions and how the MUHNAC structured and organized their collections in (Figure 3.3).

Figure 3.3- Number of collections under the Natural History MUHNAC sub-domains.

Out of all the collections within Natural History the Zoology sub-domain is the one that has the highest amount of individual collections under its area with 21 collections, followed by Paleontology with 8 collections.

A full diagram of the Natural History domain, corresponding sub-domains and individual collections is shown below (Figure 3.4).

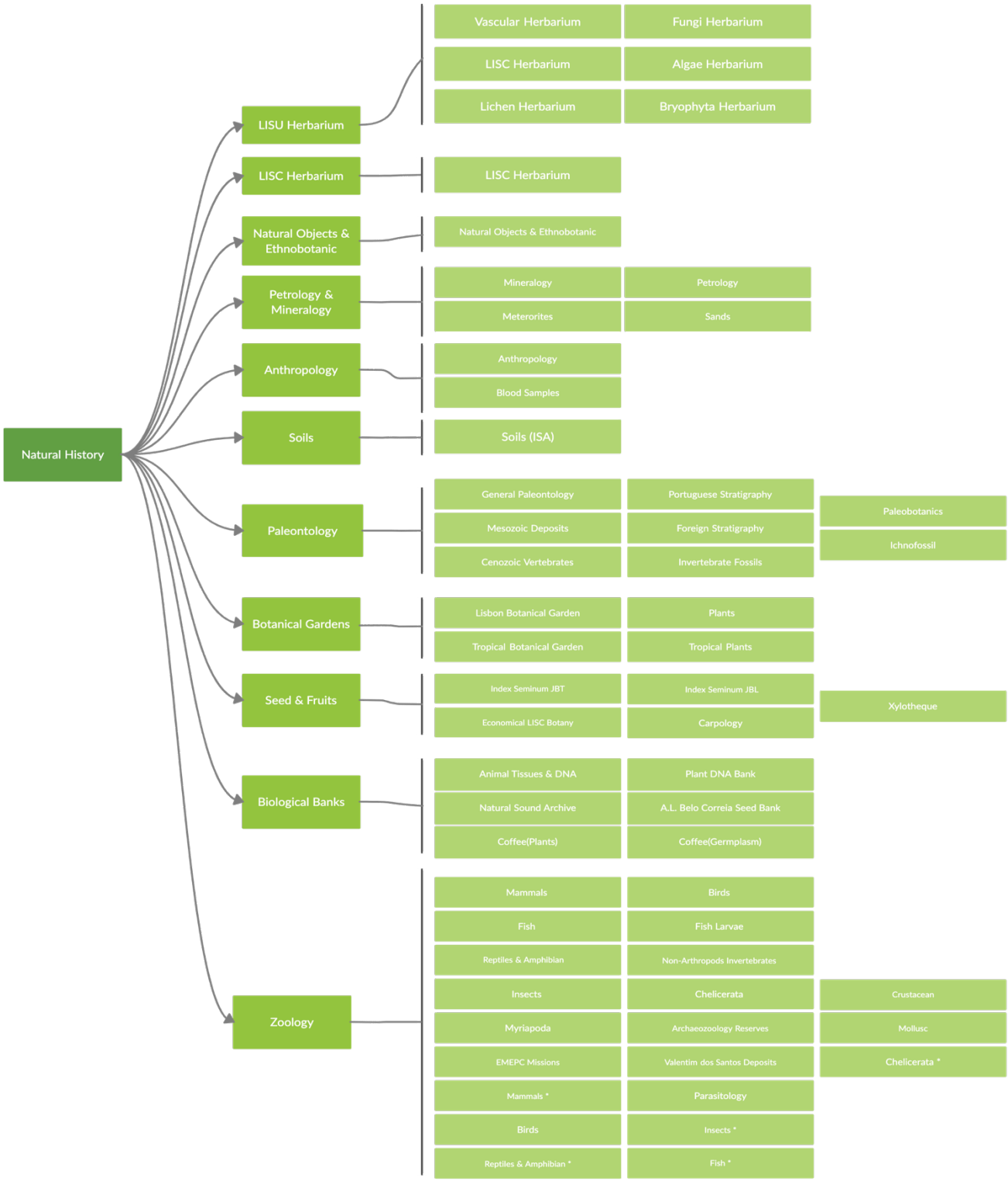

Figure 3.4- Diagram with all the divisions under the Natural History Domain and their respective collections. Diagram build from an official document provided by the MUHNAC. The * signifies that collection used to belong to former IICT but it is still considered an independent collection.

3.3 Requirements and Use Cases

In this subchapter we will focus on what are the functional requirements gathered during meetings with all the interested stakeholders as well as meeting with the teams responsible for guaranteeing the continuity and maintenance of this project.
The type of users expected to use this system, the use cases and what are the necessary steps and conditions to achieve them will also be presented in this subchapter.

3.3.1 Functional Requirements

This section presents the functional requirements (FR) found and their description.

-FR1: Unique and Specific Identifiers – All data inside the system needs to contain special and unique identifiers to unmistakably identify what is the source of that record regarding the Institution and the source of that data inside the Institution (ex: main source is MUHNAC, and inside MUNHAC belongs to a botany collection.)

- **FR2: Bulk Import of Data** – Having so much data living in different file formats such as XLSX, CSV files, Filemaker and even some third-party software's and databases, it would not be practical to have to import this data one by one, making bulk import of data from a standard file extension like CSV a necessary requirement.

- **FR3: Customized Data Export** – Museums often have collaborations in international projects based on the sharing of scientifical data (ex. GBIF, PORBIOTA, Europeana, scientific research projects), so it is important to be able to export their data and to be able to choose which metadata elements to download based on the project specific needs.

- **FR4: Image and Files Association with Records** – Collection items often have images or files with some additional information associated with them. It is required for the system to be able to allow this linkage of media to a record.

- **FR5: Roles –** The ability to create multiple roles with different access and operation permissions to ensure. Having a "curator" role with permissions only certain data and metadata elements allows for security when it comes to the data, by preventing other people to be able to access and modify the information. A "public" role would be a role with search permissions exclusively, no permissions to modify, insert or delete items.

- **FR6: Custom Search Options** – Different collections and different curators have different needs, methods and preferences when it comes to performing searches. It is important to be able to customize search forms with different searchable metadata elements and making these forms only available for the collections they were design to function.

- **FR7: Geographical Information** – Geographical information is very important as the localization of a given record is often one of the most used fields for searching and differentiation between records. Based on this, it is required for the system to able to infer this hierarchical structure based on the information specified. Ex: Lisbon is part of Portugal, if a given record has Lisbon on its geographical metadata element, when we perform a search for records in Portugal, the system needs to be able to make this relation.

Also, if given the exact coordinates where that record was obtained a visual representation in the form of a map pinpointing where the coordinates are, is desirable.

FR8: Taxonomical Information – Likewise, taxonomical information also follows a hierarchical pattern, whereas by knowing the "genus" of a biological record should be more than enough information to infer all the remaining hierarchical classifications above. It is important to provide support for taxonomy since most of the data in the MUNHAC is biology related.

- **FR9: Loans** – Museums are constantly receiving and loaning out items. Its required to have a feature that allows for the process of receiving and loaning items with all the information associated with that loan.

A summary table (Table 1) of the functional requirements is shown below.

Table 3.1- Summary listing of the functional requirements.

Functional Requirements (FR)	
FR1	Unique and Specific Identifiers
FR2	Bulk Import of Data
FR3	Customized Data Export
FR4	Image and Files Association with Records
FR5	Roles
FR6	Custom Search Options
FR7	Geographical Information
FR8	Taxonomical Information
FR9	Loans

3.3.2 Non-Functional Requirements

This section presents the non-functional requirements (NFR) and their description.

-NFR1: Open-Source Platform – It was established very early in this project that the platform used needed to be open-sourced.

-NFR2: Language of the platform – The programming language in which the platform is written was also important for continuity and future management of the project as the Central Services Department of Informatics of the University of Lisbon will be responsible for the project maintenance and if a problem emerges, they need to be able to understand the source code.

-NFR3: Storage Capacity Estimation – Understanding how the main contributors to the storage capacity increase (images, videos, files, audios) are stored for allocation of resources purposes.

3.3.3 Actors

This section describes the main actors involved in the Use Cases (UC).

- **General Public –** Represents an arbitrary person with no affiliation with the software, the Museum or the data, using the software for personal interest. For instance, a visitor interested in searching about records belonging to a specifical taxonomical group, or in a specific location.

- **Curators** – The responsible for managing the collection inside the system and their own data.

- **Admins** – The people with full access and permission inside the system, responsible to create and manage the features and requirements the curators have.

3.2.5- Use Cases

UC1 – Inserting a record Name:
Inserting a record **Actors:**
Admins, Curators **PreConditions:**

- Knowing what's the type of the record.
- Having all the information necessary, at least the mandatory information like the unique identifiers or the public accessibility.

Post-Condition:

- Record is stored within the system.
- Visibility conditional to what the curator defined, if accessibility is public, that record will appear in searches, if it is private, then only the curator can see it.
- Record is available for search. **Steps:**
  1. Select "NEW --> OBJECT".
  2. Choose the type of Object.
  3. Filling the form associated with that Object.
  4. Press Save.

UC2 – Modifying a single record
Name: Modifying a single record **Actors:** Admins,
Curators **Pre-Conditions:**

- Knowing the unique identifier for that record. **Post-Condition:**
- Record information is updated.
- Record is available for search. **Steps:**
  1. Searching that record by the identifier.
  2. Click the Edit button.
  3. Alter the desired fields.
  4. Press Save.

UC3– Deleting a single record Name:
Deleting a single record **Actors:**
Admins, Curators **Pre-Conditions:**

- Knowing the unique identifier for that record. **Post-Condition:**
- Record is deleted.
- Record no longer available for search. **Steps:**
    1. Searching that record by the identifier.
    2. Click the edit button.
    3. Pressing “Delete” on top right corner.
    4. Confirm intention to delete.

UC4 – Metadata creation
Name: Metadata creation **Actors:** Admins **Pre-Conditions:**

- Knowing the datatype of the metadata, if it’s a text-based metadata, a numerical value, coordinates, dates, currency, measurements, etc.
- Knowing the types of objects that metadata is allowed to be used.
- Knowing metadata specific parameters. **Post-Condition:**
- Metadata is ready to receive information.
- Metadata conditionally ready to be used for searches. **Steps:**
    1. Select “MANAGE --> ADMINISTRATION”.
    2. Click the “METADATA ELEMENTS” tab.
    3. Click “NEW”.
    4. Filling the metadata creation form.
    5. Press Save.

UC5– Search Form Creation Name:
Search Form Creation **Actors:** Admins
Pre-Conditions:

- Knowing what type of records the search form should be applied to.
- Metadata elements to include in search form need to check the “Can be used in search form” parameter.

Post-Condition:

- Search form is shown in the Objects it was applied to. **Steps:**
    1. Select “MANAGE --> MY SEARCH TOOLS”.
    2. Choose the target of the search form.
    3. Filling the form.
    4. Select metadata elements to be displayed in that form.
    5. Press Save.

UC6– Basic Search
Name: Basic Search
Actors: Admins, Curators, General Public **Pre-Conditions:**

- Knowing the ID or the Preferred Label of the record to search. **Post-Condition:**
-

Not Applicable **Steps:**

1. Select “FIND --> OBJECTS”.
2. Choose “BASIC SEARCH”.
3. Write ID or Preferred Label in Search Box.
4. Press Search.

UC7– Advanced Search
Name: Advanced Search
Actors: Admins, Curators, General Public **Pre-Conditions:**

- Field desired to search needs to be included in a search form.

Post-Condition: •
Not Applicable **Steps:**

1. Select “FIND --> OBJECTS”.
2. Choose “ADVANCED SEARCH”.
3. Choose a search form on right top corner.
4. Write the search term on one of the form fields or multiple search terms in the multiple fields.
5. Press Search.

UC8– Search via Browse Name: Search via
Browse
Actors: Admins, Curators, General Public **Pre-Conditions:**

- Not Applicable **Post-**

Condition: • Not Applicable **Steps:**

1. Select “FIND --> OBJECTS”.
2. Choose “BROWSE”.
3. Click on one of the boxes on screen.

UC9– Data Import
Name: Data Import
Actors: Admins, Curators **Pre-Conditions:**

- Mapping file for that data needs to exist in the system.
- Files can’t contain macros. **Post-Condition:**
- Data is now part of the system.
- Records are searchable. **Steps:**
    1. Select “IMPORT --> DATA”.
    2. Select the appropriate mapping file.
    3. Upload Source Data file.
    4. Press “Execute Data Import”.

•

UC10– Data Export
Name: Data Export
Actors: Admins, Curators **Pre-Conditions:**

Not Applicable **Post-Condition:**

- Data is downloaded to your computer in the desired format. **Steps:**
  1. Perform any type of Search (Basic, Advanced or Browse).
  2. Select the "Export Tools" icon on top right corner.
  3. Choose Download format (TSV, CSV, WORD, PDF, XLSX).
  4. Click the "→" icon.

UC11– Image Association Name:
Image Association **Actors:**
Admins
Pre-Conditions:

- Image needs to have the same name as the unique identifier for the record it belongs.
- Images to import need to be at: webserver_root/providence/import **Post-Condition:**
- Record now display the image when searched.
- Possibility to Browse only records containing images. **Steps:**
  1. Select "IMPORT --> MEDIA".
  2. Fill the form with the import definitions and settings.
  3. Press the "Execute media import" option.

UC12– Batch Editing Records Name:
Batch Editing Records **Actors:**
Admins, Curators **Pre-Conditions:**

- Identify all fields that need correction.
- Knowing the correct replacement information. **Post-Condition:**
- All records are updated to contain the new information.

Steps:
1. Perform a search that encompasses all records with wrong information.
2. Click "Set Tools" on top left corner.
3. Select "Create a Set from Results" option.
4. Go to "MANAGE --> MY SETS --> ALL SETS".
5. Identify the set just created.
6. Press the "Wand" symbol under "#Items".
7. Select the field(s) you'd like to have the information changed.
8. Fill the text box with the new information.
9. Click "Execute batch edit" on top of the page.
10. Confirm intention to change the records.

UC13– Loan Creation Name:

-

Loan Creation **Actors:** Admins,
Curators **PreConditions:**

- In case of a Loan-out, the record to loan must be in the system.
- Need to know the unique identifiers of records to loan. **Post-Condition:**

  Loaned records can be seen when performing loan searches **Steps:**

  1. Select “NEW --> LOAN --> Type of Loan” 2. Fill the initial form with loan related information.

-

3. Press Save.
4. Select the “Relationship” tab on the left of the page.
5. Add the Objects to loan using their ID.
6. Press Save

UC14– Custom Interfaces

Name: Custom Interfaces **Actors:**
Admins
Pre-Conditions:

- Metadata to display in interfaces needs to check the “Can be use in display” parameter.
- Metadata to display needs to be linked to the same type of the custom interface. **Post-Condition:**
- When inserting/modifying/searching for a record, if an interface for that Object type exists, it will be displayed over the default CA interface.

Steps:

1. Go to “MANAGE --> ADMINISTRATION”.
2. Select the “User Interfaces” tab on the left of the page.
3. Choose a Type for the interface.
4. Press the “+” icon
5. Fill the form with interface related fields.
6. On the “Screens” field, add screens to group related metadata elements.
7. Press Save.
8. Go back to the “Screens” field and click the edit button.
9. Drag the metadata element from the available list to the list of metadata elements to show in that screen.
10. Press Save.
11. Repeat step 8 for all the screens.

UC15– Permission Assignments

Name: Permission Assignments
Actors: Admins, Curators (in case they have permissions to give permissions) **Pre-Conditions:**

- Curators need to have a valid CA account. **Post-Condition:**
- Curators have permissions to realize the operations they were given permissions to. **Steps:**
  1. Go to “MANAGE --> ACCESS CONTROL” 2. Select the “USER LOGINS” tab.
  3. Click the “Edit” icon on the user to give permission.
  4. Assign a role to that user.
  5. Press Save.

Observations:

Roles have a predetermined set of permissions. If a particular user doesn’t fit any of the existent roles, on the “ACCESS CONTROL” menu select the “ACCESS ROLES” tab on the left of the page and select the “New role” icon. Give it a name and identifier and choose the permissions this role should have from the list of permissions. Press Save, go back to the “USER LOGINS” tab and select the user again assigning him the newly created role with the custom permissions.

In Table 2 a summary of all the Use Cases and the Actors involved in them is presented.

Table 3.2- Listing of the use cases and the actors allowed to participate in them.

	Admin	Curators	General Public
UC1 – Inserting a record	x	x	
UC2– Modifying a single record	x	x	
UC3– Deleting a single record	x	x	
UC4– Metadata creation	x		
UC5– Search Form Creation	x		
UC6– Basic Search	x	x	x
UC7– Advanced Search	x	x	x
UC8– Search via Browse	x	x	x
UC9– Data Import	x	x	
UC10– Data Export	x	x	
UC11– Image Association	x		
UC12– Batch Editing Records	x	x	
UC13– Loan Creation	x	x	
UC14– Custom Interfaces	x		
UC15–Permission Assignments	x		

3.4 Gap analysis

This section will cover the initial idea behind the process of choosing the software used throughout this project, providing a comparison between multiple collection management software's and the reasons behind the final decision.

3.4.1 Introduction

To fulfil the goals of this project in adequate time and investment, it was decided to investigate available open-source software that covered some of the key requirements of the MUHNAC and if those software were also extensible in order to support the implementation other requirements if needed. Specify, Collective Access and CollectionSpace were the three software that made cut for the final analysis based on the information provided by their documentation and the functionalities they provided by default.

3.4.2 Specify

Specify was one of the open-source software analyzed during the state of the art as a possible contender to be the chosen platform in which this project would be implemented. Specify also had the advantage of already being used by some curators inside MUNHAC for their collections providing some end-user experience and feedbacks.

Specify is an open-source biological collections management software which can be helpful since the MUHNAC has a high amount of biological collections, nonetheless, the whole idea behind this project was to gather all collections regardless of their domain in a single collection management software. Nevertheless, attempts were made to install a local instance of Specify to get an idea of how this software worked, its capabilities and if it would be easy to extend and implement new features, specifically support for non-biological collections and their requirements.
Unfortunately, the official documentation had some flaws and the instructions on the installation manual weren't clear and detailed. Several attempts were made to contact the Specify support via their official support e-mail and forum on their website, but no response was ever obtained. The lack of responses and the known fact that non-biological collections were not supported by Specify lead to a shift in direction, and other software had to be considered as a possibility.

3.4.3 Software Comparison

After the initial Specify attempt it was made clear there was a need to search other software choices for this project. In one of the meetings with the stakeholders, two more collection management software apart from Specify were mentioned as possible candidates for this project: "CollectiveAccess" and "CollectionSpace".

In Figure 3.5 is presented a comparison between these software. Only some of the broader overall features were considered since at the beginning of the project development the listing of the functional requirements was not established yet.

	Specify7	CollectiveAcess	CollectionSpace
Open-source	Yes	Yes	Yes
Ontologies	N/A- its not described. Mostly used for biological data, can assume it follows darwin-core ontology, but doesn't support non-biological data.	Dublic-Core Darwin Core - Biological Samples EBU-core - Audio and video resources PBCore- Audio and Video CDWA- Art and Material Culture ISAD(G)- Archives, Libraries, Museumns, Manuscripts VRAcore - Visual Culture and Images for it	Documentation says it supports art museums, historical society, harbaria and botanical gardens, no specific mentions of the ontologies' names.
Language	Python(Django)	PHP	Java
Extensible	Possibly yes, but very hard to understand code structure, over 100 files with close to no comments in code.	Based on github yes. Understandable code structure and file system.	Documentation says it can be extended. Didn't look at the code.
Import Feature	Yes, but no mention of input formats in documentation	Yes. Formats: XLSX, TSV, CSV, XML, MARC, Filemaker.	Yes. Formats: XML, MARC, Dublin Core, XLSX.
Export Feature	Yes, but no mention of output formats in documentation	Yes. Formats: XLSX, TSV, CSV, DOCX, XML, PDF.	Yes. Formats: E-mail, PDF, XML, XLSX.
Loan Support	Yes	Yes	Yes
Feature Link	https://github.com/specify/specify7	collectiveaccess.org/features	collectionspace.atlassian.net/wiki/spaces/COL/overview collectionspace.org/faq/#userfriendly
User Differences	Yes	Yes	Yes
User Roles	Yes	Yes	Yes
Images, Videos, Audios storage	Cant tell, theres a column named " card image full path" which describes the full path, and theres a collumn called "card image data", type "memo" with lenght of 16.000.000	Stores images, videos, audios in folders and has reference to path in database.	Stores images, videos, audios in folders and has reference to path in database.

Figure 3.5- Comparison between Specify, CollectiveAccess and CollectionSpace on some overall features and characteristics.

When it comes to being an open-source software all 3 different systems met this requirement. Ontologies represented one of the most important features in this comparison. It is of extreme importance for the MUHNAC that their data is described using international nomenclature, for the purpose of collaborations, both international and national. Specify did not provide names of specific ontologies, but some of the MUHNAC stakeholders familiarized with the software stated that is main use is for biological data, and the main ontology used there is the Darwin Core Ontology. CollectionSpace's documentation stated their platform supported art museums, herbaria and botanical gardens, but with no mention of the actual ontologies' names. CollectiveAccess provided a full list of ontologies it supported by default, including Darwin Core for biological data, EBU Core for Audio and Video Resources, CDWA for Art and Material Culture, ISAD(G) for Archives, Libraries, Manuscripts, between others. The language in which these systems were written was also considered as it was part of the nonfunctional requirements. When this project is finished if a decision between the stakeholders is to implement this system and get it to the production stage, ready to be used by the MUHNAC, it is going to be maintained by the Department of Informatics of the Central Services of the University of Lisbon, a language they support and can understand is required in case there is the need to fix some problem. Import and Export features were also part of the functional requirements. Import is essential due to the high amount of data the MUHNAC has; it wouldn't be practical to have to insert hundreds of thousands of records manually one-by-one. Export because the museums have external collaborations with other international/national projects like GBIF and PORBIOTA and they need to submit parts of their data. User difference/roles was also part of the functional requirements. With so many different types of collections and each collection with one or more curators, is recommended that each curator can insert/modify or delete records from the collections they are responsible for. Having the capability to assign certain permissions to specific users regarding the different types of data is mandatory. Lastly, on a more technical note, we are living in a world that is more digital by the day and following this trend the MUHNAC has been making an effort to digitalize their records. For the purpose of server infrastructure and technical specificities, it was important to understand how the images, videos, audios were stored, as they are the main contributors for the increase of storage capacity required.

3.4.4 Decision Making

Considering the outcome of the gap analysis and the functional requirements for this project,

Collective Access (http://collectiveaccess.org, developed and maintained by Whirl-i-Gig, Release 1.8) shown to be the most likely to have better success as the chosen software for managing the collections data as it provided by default support for the following requirements:

- Support for all data, biological or not.
- Built in support for the Darwin-Core Ontology.
- Access to bulk import and export of data from/to a variety of different file extensions.
- Differentiation between user and the operations they can perform.
- Familiar programming language.

Additionally, a similar project was recently implemented using the same software for the “Plataforma dos Açores Digital” providing a model, functional system and a guarantee of the system capabilities.

3.5 Architecture

This section will cover the general architecture behind CollectiveAccess and how to understand the underlying structure that makes the software work as intended. It will also cover the methodology and steps taken from receiving a dataset to the import of it into the system.

3.5.1 Database Structure

After the installation process is completed, we have access to the database and all the tables CA creates. On a first database look, it may seem complex to comprehend exactly how this software works, especially when looking at the results of querying the database for the number of tables created:

"""SELECT count(*) AS TOTALNUMBEROFTABLES
FROM INFORMATION_SCHEMA.TABLES
WHERE TABLE_SCHEMA = 'test' """

"TOTALNUMBEROFTABLES = 224"

To work effectively with the software, it is critical to understand the fundamental components of a CollectiveAccess database. While CollectiveAccess provides great flexibility in terms of the specifics of the data model, in terms of defining our own fields, relationships and constraints, the general structure is fixed. CollectiveAccess defines fourteen types of categories that cover all the possibilities for our data to exists in. These are referred throughout the documentation as "Primary Types" or "Basic Tables” and they are the following:

- Lots (ca_object_lots)
- Objects (ca_objects)
- Entities (ca_entities)
- Collections (ca_collections)
- Occurrences (ca_occurrences)
- Loans (ca_loans)
- Places (ca_places)
- Movements (ca_movements)

- Sets (ca_sets)
- Set Items (ca_set_items)
- Representations (ca_object_representations)
- Storage Locations (ca_storage_locations)
- Lists (ca_lists)
- List Items (ca_list_items)

Besides the primary tables mentioned above, a huge portion of the total tables present in the database are the relational tables. Every basic table establishes a relation with the other basic tables. Considering the Objects table (ca_objects) as an example, the relations with the other tables are represented as follows:

ca_objects_x_collections ca_objects_x_entities
ca_objects_x_objects
ca_objects_x_object_representations
ca_objects_x_occurrences ca_objects_x_places
ca_objects_x_storage_locations ca_objects_x_vocabulary_terms
ca_objects_labels

This "basicTable_x_basicTable" pattern repeats for every single one of the tables mentioned above and it's how different data within the system is able to relate and be referenced with a particular record, for instance one "Object" record will be stored in the "ca_objects" table, if that record is part of a collection, that collection will be defined in "ca_collections". The "ca_objects_x_collections" will store both the id of the object and the id of the collection it belongs to.

The remaining tables present in the database are mostly system necessary tables for permission assignments, interfaces and their metadata, custom forms amongst other features of CollectiveAccess.

3.5.2 Intrinsic Fields

Although CollectiveAccess gets most of its praise for being a flexible software, when it comes to the data types it supports and the metadata its users can create and associate with their data as they need to, in order to properly function, there is a hardcoded set of attributes every record needs to have based on their Primary Type, these attributes are referred to as "Intrinsic Bundles".
Intrinsic bundles are the CollectiveAccess way of ensuring the correct behavior of the system, and they are the only hardcoded set of attributes each record inside the system must have and they vary between the different Primary Types.
Not all the intrinsic fields are required to be filled by the user with the exception of the "idno" field, which is the unique identifier for every record. The remaining ones if no information is provided, CA will set them to their default value.
In Table 3 is showed what are the intrinsic fields and their description that a record of type "Object" has by default.

Table 3.3- Intrinsic fields of an Object and its Description

Bundle **Name**	**Description**

idno	Object identifier field (often used for current accession number).
locale_id	Object record locale drop-down
item_status_id	Object status value based upon values in object_statuses list
acquisition_type_id	Object acquisition value based upon values object_acq_types list
source_id	Object source value based upon values object_sources list
extent	Numeric extent
extent_units	Units of extent
access	Control of public access using values defined in the access_statuses list
status	Indication of current workflow status as defined in the workflow_statuses list
ca_objects_deaccession	Deaccession status of an object.

The “access” intrinsic field is one of the most relevant from this list. Every time a search is performed, CollectiveAccess checks what is the value assigned to the “access” metadata. In this field our options are “accessible to public” and “not accessible to public”. Should there be a record with sensible information, setting the “access” element to “not accessible to public” will remove that record for search results, making it private an only accessible with permissions.

3.5.3 General Methodology

The methodology proposed in this project can be applied to nearly every dataset. Its main architecture is represented in detail in Figure 3.6 followed by a description of the different stages.

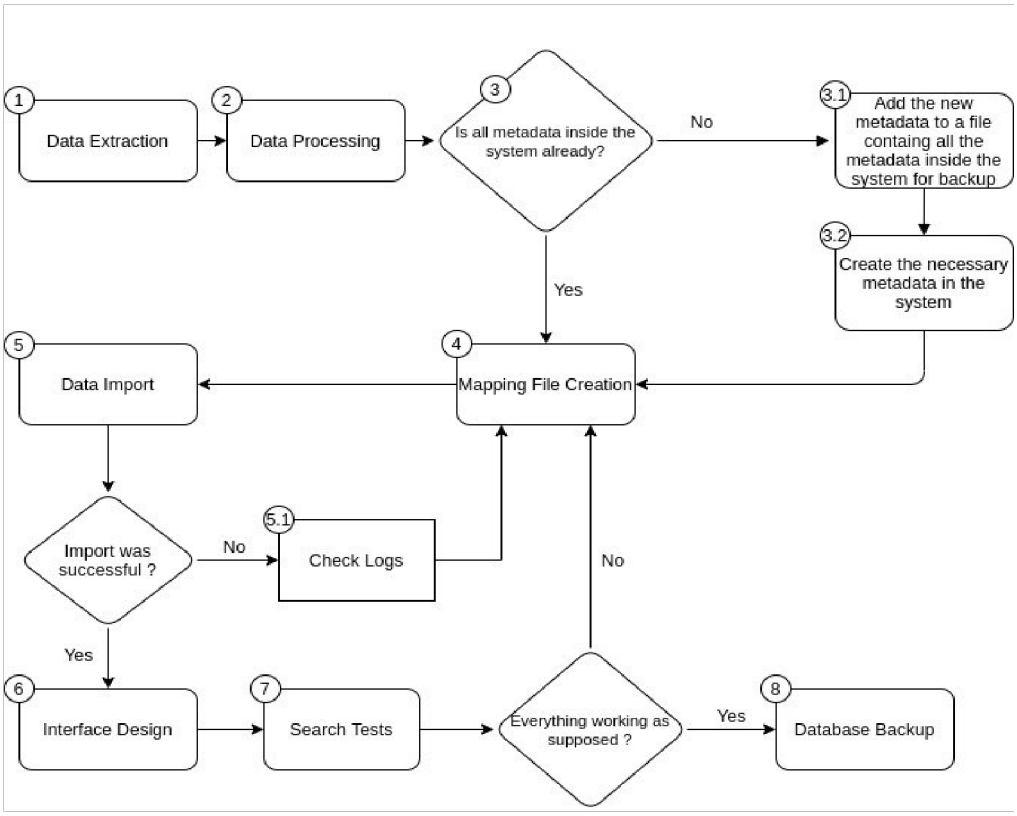

Figure 3.6- Workflow of processes diagramming the methodology starting with the data extraction (1) all the way to the backups (8).

1- **Data Extraction:** Extracting the data from its original source, whether it is in a database, a thirdparty software or a data file.

2- **Data Processing:** Making sure all data follows the Unicode Standard, correction of typing mistakes, replacement of non-allowed characters.

3- **Metadata Verification:** Comparison between the metadata in the source data against the metadata present inside the system to try to find matches. If metadata is not present in the system yet:

3.1- Add this new metadata to a helper file to keep track of all existent metadata in the system and for later taking advantage of installation profiles to automate metadata creation.

3.2- Create the missing metadata elements in the system.

4- **Mapping file Creation:** Creation of a mapping file to establish the relations between the source data and the system.

5- **Data Import:** Using the mapping file created previously to import the data to the system.

5.1- If the import fails, a look at the logs is a good way to understand why. Logs can have multiple settings, like debug, error or information and most often points us to what is wrong with the mapping file.

6- **Interface Design:** Creation of the custom interface with the information specific for the dataset just imported.

7- **Search Tests:** Performing search test on multiple fields of the imported dataset to make sure everything is working correctly.

8- **Database Backup:** Making a dump of the database, to ensure the current implementation and data is safe.

4 Implementation

This section will cover some of the functionalities implemented throughout this project, starting with the installation of the software, explaining the datasets used for testing, the process behind handling large numbers of metadata elements and making sure they follow the Darwin-Core ontology, the process of importing the data into the system and some specific and more challenging implementations like handling taxonomical or geographical information. Some technical details are presented with the goal of providing a record for the future development and management of the system.

4.1 Installation

To install CollectiveAccess the first step is to make sure all the requirements for the software to run are fullfilled. Being a web-based application written in PHP, the first piece of software required is a webserver with PHP support (Apache 2.2 or 2.4 is recommended). CA utilizes MySQL as its relational database management system, and it is recommended to use either one of the 5.5, 5.6, 5.7 versions. Version 8 of MySQL is confirmed to not work properly, other versions should work fine as long as they support InnoDB tables. One more piece is missing and that is the actual PHP programming language, the documentation states that its strongly recommended to use PHP 5.6 or better, and the following PHP extensions: XIp, cURL, libXM, mbstring, iconv, EXIF, JSON, MySQL, posix and OpenSSL or mcrypt. These 3 requirements (Webserver, MySQL and PHP) were met taking advantage of the LAMPP web service stack which provided all the software necessary for running this application.
After everything is set up, an empty MySQL database should be created, given a name and a login for it with full read/write access and the GitHub repository containing the CollectiveAccess (https://github.com/collectiveaccess/providence) should be cloned into the root of the webserver instance. The settings on the setup file need to be changed in order to reflect the database and login information, followed by triggering the installation process, selecting one of the pre-defined installation profiles, and the software should be launched on the web-browser.

4.2 Datasets

African Decapoda

The first dataset processed was the African Decapoda Collection, curated by Dr. Alexandra Cartaxana. This dataset provided 1118 records described by 113 different metadata elements organized within eight groups:

-Identifiers: This group contained metadata elements like “InstitutionCode”, “CatalogNumber”, “CataloguedBy”, “dateCatalogued”, amongst other elements used to provide the general information about the records identifications and who was responsible for its cataloguing. **Location:** Information about where this record is stored under the MUHNAC building. Metadata elements like “disposition”, “locationMuseumSpecimen”, “locationMuseumTissue” are present here.

-Taxonomical Information: Taxonomical related metadata containing the records taxonomical classification, taxon rank, previous identifications, and taxonomic status is found in this group. **-Event:**

Information on the event that led to the gathering of the records, dates and times when the event started and ended, sampling protocol used, sampling observations, and field notes are some examples of annotations that are part of this group.

-Geographical Information: This group contains the most metadata elements out of all with 31 elements. Information about the geolocalization of the record is here, including elements like "continent" to "country" passing through coordinates systems, both in decimal and Verbatim forms, and ending on the georeference elements like "georeferencedDate", "georeferencedProtocol", "georeferenceRemarks".

-Collection Specific Elements: Information like the sex of the records, the lifeStage, number count, length, reproductive conditions as well as measurements like "cefalotoraxlegth (mm)", "carapace length (mm)", "carapace width (mm)", "weight" form this group.

-Preparations and Media: This last two groups contained information about the preparations in which the record was stored, if it was a record obtained through a donation and general information about the media associated with the records if it exists, and general notes about the condition of the record.

Overall, out of the 113 metadata elements 74 elements could be describe using the terms defined by the Darwin Core Ontology and provided a solid mix between textual based metadata, numerical metadata and coordinates-like metadata that were experimented on and implemented for future datasets.

LISU Herbarium

This dataset was provided by Dr. Judite Alves and contained the vascular plants collections of LISU Herbarium. It is a smaller dataset when compared to the Decapoda collection with only 206 records and 81 metadata elements.
This dataset had the particularity of being one of the collections that was previously stored on Specify and as a result of that when the data was exported to a XLSX format due to the database structure of Specify, a large number of elements became duplicated (e.g., "family1"," family2", "family3"). Most of the metadata was common with the previous dataset with changes almost always occurring in the collection-specific elements.
Nonetheless, the metadata elements were extracted, analyzed and prepared for their automatic creation, but a follow-up meeting with Dr. Alexandra Cartaxana regarding this dataset particularities resulted in the decision of putting this collection on hold while a decision on how to import this dataset, whether the original formatting with the repeated elements or the separation of them into new records was being discussed internally. **LISC Herbarium**

Finally, the ultimate dataset that served as the demonstration of the capabilities of the system, was the LISC Herbarium collection curated by Professor Maria Cristina Duarte, co-supervisor of this project.

With 63562 records this dataset is, by far, the largest received but only having 22 metadata elements, which represented a partial amount of all the metadata available for this collection (more than a hundred). Notwithstanding this issue, this dataset provides the necessary elements to present a demonstration of all the implemented functionalities and their behavior to the stakeholders, conduct the usability test on both curators and public.

4.3 Data Processing and Metamodels

The first step upon after receiving a dataset is to make this data go through a processing stage. This processing stage was implemented using python and it is responsible for handling various operations, ranging from spelling mistakes, correction of non-Unicode and UTF-8 values and extracting the metadata elements for that particular dataset.

The extracted metadata elements were analyzed to ensure they were compliant with the Darwin-Core ontology and, if they were not, if there was a Darwin Core term that could be applied to describe the same property of that particular element. For that purpose, every metadata element was compared with the list of terms described by the ontology to verify if there was a correspondence between them.

For instance, a matching example would be something like shown in Figure 4.1, where the first column contains all the positive matches with Darwin Core alongside the formal definition for that term.

Fish Collection	Definition
collectionCode	The name, acronym, coden or initialism identifying the collection or data set from which the record was derived
scientificNameAuthorship	The authorship information for the scientific name formatted according to the conventions of the applicable Nomenclatural Code

Figure 4.1- Example of a positive match between a metadata element with a Darwin-Core term alongside its official description.

Some metadata elements were collection-exclusive and described specific properties of that data without any relation to Darwin Core (Figure 4.2). The definition and meaning of those elements had to be inferred by the contents of the data or if the metadata name is self-explanatory.

Fish Collection	Definition
locationMuseumSpecimen	The location of the item in the Museum

Figure 4.2 -Example of a collection-specific metadata element with the inferred description

After every match or non-match an auxiliary file containing all the different metadata was produced to help speed up the process of verifying redundant metadata and to help track every metadata necessary to be created in the system. This file was later used in the metadata automatization process.

4.4 Data Organization

After the software is installed and ready for use and before starting to introduce new records to the system, it is essential to have a defined idea on how this data is organized. It is through this organization that custom interfaces assignments custom access permissions and custom metadata is ensured.
For this project the structure of the data followed the organization provided by the MUHNAC and to make all records belong to a single point in the structure, CollectiveAccess's list were used to create this organization (Figure 4.3).

Figure 4.3- Representation of the internal MUHNAC structure using CollectiveAccess' list.

After the full list is created, its structure is reflected on the "New -> Object" menu (Figure 4.4).

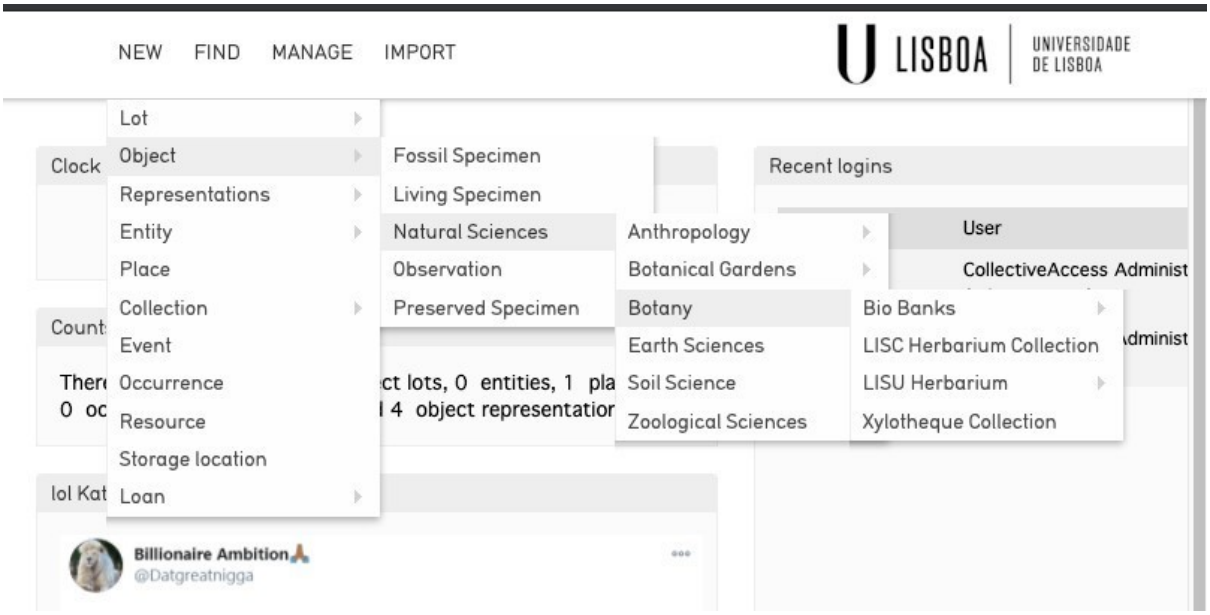

Figure 4.4- Visual representation of the internal MUHNAC structure on the data insertion menu.

Because collections are the containers of the data, and every collection is independent and perceived as being its own type, some elements used in the creation of the data organization have a purely structural role . Nodes that do not end with the word "Collection" are the structural nodes used to organize the collections into their domains and they are not clickable and available for data insertion. Every collection has its own place in the ends of this hierarchical structure. These structural nodes not only serve for organizing the collections, but also allow for grouped searchs. It is certainly possible to search for all data with a specific type, getting us all records inside that collection, but because this feature was implemented using a hierarchical list, the end nodes inherit their parents node, so a search by the "Botany" type would give us all records belonging to collections further down the hierarchy.

4.5 Mappings

In CollectiveAccess there are two different ways of introducing new records to the system, it can either be done manually by selecting the Object type under the “NEW” menu on the navigation bar (Figure 4.5) or choosing the “IMPORT” menu.

Figure 4.5- Navigation bar for the web portal.

4.5.1 Simple Mapping

To use the import feature and keeping in mind the versatility of the software, in terms of metadata creation, assigning specific metadata to specific collections and different collections topologies, there is the need to generate an intermediary file (mapping file) that establishes the connections and relations that our data source has with the information inside the CollectiveAccess.
The creation of a mapping file has to be individualized to a specific data source, meaning there is the need to create a mapping file for every different data source before its import. This mapping file also needs to follow a predetermined set of rules and constraints defined by CollectiveAccess.

A simple mapping example is presented in Figure 4.6, corresponding to the collection source data, and Figure 4.7, related to the mapping file.

exportcristinaId	Catalog_Number	Cataloged_Date	Determiner	Family	Genus	Full_Name
4	LISC022913	2009-4-29	A.F.Martins	Simaroubaceae	Brucea	Brucea antidysenterica
5	LISC022914	2009-4-29	I.Nolde	Simaroubaceae	Brucea	Brucea antidysenterica
6	LISC022915	2009-4-30		Simaroubaceae	Hannoa	Hannoa chlorantha
7	LISC022916	2009-4-30		Simaroubaceae	Hannoa	Hannoa chlorantha

Figure 4.6- Partial view of the source data from the LISC Herbarium Collection

Rule type	Source	CA table.element
Mapping	2	ca_objects.idno
SKIP	1	
Mapping	2	ca_objects.catalogNumber
Mapping	3	ca_objects.dateCatalogued
Mapping	4	ca_objects.determiner

Figure 4.7- Partial view of the mapping file created for the LISC Herbarium Collection

In the first column (Figure 4.7) rule types are defined. They can be either “Mapping”, “Skip”, “Constant” or others. The “Source” column references the column number in the imported source data that is going to be used accordingly to the rule type, and in the “CA table.element” column the relation of source data with metadata inside CollectiveAccess is establish. This particular column always needs to follow the “TableName.metadataElementID” format. These three columns are mandatory and are the basis of the mapping.
So, analyzing the first two columns of Figure 4.7, what is being done in the first row is establishing a mapping relation (Rule type) between the second column of the source data (Catalog_Number), which represents a unique identifier for every record and the metadata element “idno” from the “ca_objects” table, which is the Objects table intrinsic field for unique identifiers. In the second row we're saying to

CollectiveAccess to skip the first column of my source data (might be a blank column or a column we do not want to import) in this case is a metadata element we do not want to import, as is the case.

The same column of the source data (Catalog_number) is also being used to populate a metadata element with the same name, in the third row of the mapping file (Figure 4.7).
After this pattern is repeated for every metadata in our source data, some additional setting (Figure 4.8), also need to be filled.

	Setting name	Setting value
Setting	name	import_cristina_LISC
Setting	code	LISC_import
Setting	inputFormats	XLSX
Setting	table	ca_objects
Setting	type	MachineObservation
Setting	numInitialRowsToSkip	1
Setting	existingRecordPolicy	none
Setting	errorPolicy	ignore
Setting	archiveMapping	yes
Setting	archiveDataSets	yes
Setting	basePath	#XML tree#
Setting	locale	DEFAULT

Figure 4.8- Partial view of the additional settings for the mapping file created for the LISC Herbarium Collection. Every mapping file needs to have a name and a unique identifier (Setting name = name and code), and file extension of our source data needs to be typed, in this case is a XLSX file. Very importantly, the main table our data is going to be inserted in needs to be specified and can only be one table (Setting name = table). If the data requires the combination or mapping from different tables, additional parameters in the mapping file have to be employed. Besides these settings there are a couple of other settings to define a number of rows to skip, in case of headers in our data, error handling settings, among others

4.5.2 Refineries

Every mapping file can ultimately only have one primary table as its destination, but often times is required to use values belonging to different tables to populate certain metadata elements or to establish relations. Refineries are one of the optional settings that can be present in the mapping file and it is through the use of them that the relations between different tables are establish in the mapping files. One example on the use of refineries in this project is shown in Figure 4.9.

Rule type	Source	CA table.element	Refinery	Refinery parameters
Mapping	6	ca_objects.taxonomy	listItemSplitter	{"list":"test_taxonomy", "dontCreate":"1", "ignoreParent": "1", "matchOn": ["labels"]}

Figure 4.9- Example of the use of a refinery in a mapping file.

In this example the sixth column of Figure 4.6 has information regarding the taxonomical classification of the records, in particular the genus. Since the taxonomical information was implemented as a controlled vocabulary using CollectiveAccess lists where every taxonomical classification would be a list item, when there's a need to map the values present in the data source to the values present in the list

containing the taxonomy, a refinery needs to be employed to make these relations. In this case, particularly the objective is trying to map to list items so a "listItemSplitter" is used.

When using a refinery there is the need to also use the "Refinery parameters" setting and based on the type of refinery and primary table associated with its different parameters are available to use.

These parameters are passed through using a JSON format (Figure 4.9, Refinery Parameters column). In this case, the first parameter is called "list". This is a mandatory parameter as it tells CA what is the list it should use to try to perform matches on. The default behavior for this particular type of refinery is to create a new list item in the list if no match is found. This can be particular troublesome due to spelling mistakes, since a huge quantity of redundant list items with slightly different spellings will be “obtained”. This behavior is overwritten with the "dontCreate" refinery parameter set to 1.

Other default behavior that needs to be overwritten is the fact that most lists used for controlled vocabulary contain only one "node"; there is not the notion of a hierarchy where a node has different nodes inside it and each of those nodes also have nodes inside them, so if the "ignoreParent" parameter is not present or set to 1, CA will only look for matches in the first node.

The final parameter is the "matchOn". Every list item, besides its unique identifier, also contains labels (the text to display on the screen), this parameter tells CA in what property of the list item metadata elements it should try find a match.

Overall, the process of creating a mapping file can be as complex as your source data requires it to be with multiple options to choose from and even more parameters for every option.

There is not a "universal" mapping file that works for every source data, as every system is different with different metadata and relations. When considering that the MUHNAC has 97 different collections, if this number was roughly translated into 97 different data sources there would be the need to create 97 different mapping files, and each of them would only work for the source data it was designed to. Any changes on the source data structure would implicate the complete re-creation of the mapping file.

4.6 Taxonomy

4.6.1 Introduction

Although not every data in MUHNAC falls into a Biological domain, there is still a huge portion that does. Since CollectiveAccess is so flexible when it comes to data types it cannot predict out of the box every situation and requirement its users will need and unfortunately pre-built support for taxonomical classification is something that might have been considered as a TODO feature for Collective Access as the majority of museums have some sort of biological data.
In the documentation a "taxonomy" datatype, which would be connect to the uBios and ITIS API and provide an auto-complete feature, is described, but unfortunately, according to CA's official forum, that service has been discontinued.
In order to implement this feature in CA, a solution utilizing the built-in functionalities had to be developed so the relation of a specific metadata element with a group of items could be used as a controlled vocabulary to populate that metadata element.

Is also needed, for CA, to be able to infer all the information that is not explicitly present, since taxonomical classification follows a hierarchical structure and in principle knowing the genus of a record is enough information to complete the remaining taxonomical classification above in the hierarchy.

4.6.2 Implementation

To achieve the correct behavior expected when handling taxonomical information CA's list were once again the feature used.
The first step was to have a file containing the taxonomy. That file was obtained using Catalog of Life export feature [13] as it provides the ability to export a CSV file containing all the taxonomical classifications from the kingdom to the infra-specific epithet. Catalog of life was chosen as the portal to provide the taxonomy, as some of the web-portals, specifically the Naturalis, used the same taxonomical classifications obtained from Catalog of Life.

After obtaining the file, and because introducing all the different taxonomical classifications by hand would not be practical and take way too long, there was the need to decide between writing a script that analyses that file, processes it and does direct inserts in the database or take advantage of CA's installation profiles. The installation profile is an XML file that CA uses with some general metadata and lists, just so when the software is installed it is not a completely empty system.

The decision leaned towards the installation profile route, for several reasons, including the possibility of adding extra information like metadata or extra needed lists to it and having them being created automatically as soon as the software installs serving as a second source of backups outside of the database dumps. The XML format for list items is as follows:

```
<lists>
 <list code="Taxonomy_plantae" hierarchical="1" system="0" vocabulary="1">  <labels>
<label locale="en_US">
    <name>Taxonomia_plantae</name>
  </label>
 </labels>
  <items>
   <item default="0" enabled="1" idno="Plantae">
<labels>
     <label locale="en_US" preferred="1">
      <name_singular>Plantae</name_singular>
      <name_plural>Plantae</name_plural>
     </label>
    </labels>
<items>
      <item default="0" enabled="1" idno="Tracheophyta_phylum_plantae">
       <labels>
        <label locale="en_US" preferred="1">
         <name_singular>Tracheophyta</name_singular>
```

```
        <name_plural>Tracheophyta</name_plural>
       </label>
      </labels>
     </item>
    </items>
   </item>
  </items>
 </list>
</lists>
```

After the opening <lists> tag, which indicates this section of the XML file is were all list should be, the <list> indicates the beginning of a new list whereas its "code" attribute indicates the unique identifier for this list, the "hierarchical" attribute set to "1" indicates this is going to be a hierarchical list. The "system" attribute differentiates between CollectiveAccess own lists and lists generated by the users. In this case it was a list created within the project so the attribute has the value "0", and the "vocabulary" attribute indicates weather this list is going to be used as a controlled vocabulary or not, meaning that if a metadata element is connected to this list, only values in this list will be accepted and valid.
The <name> under <label> is just the name to be displayed on the screen.
The <items> tag is where the real items from that list will be, as the previous tags only defined an empty list. Inside it, is the place for the individual items under the <item> tag, and each item will have a "default" value which means that if no value is passed for this list that item should be the one that populates the metadata element, and there can only be 1 default item per list. They have an "enabled" attribute; meaning this particular list item is available to be used and an id represented by the "idno" attribute. Afterwards, is the definition of the labels; they are just the textual values of the item and they are what is going to be shown in the metadata. Lastly, there is have <items> tag meaning this item will have sub items, creating the idea of a hierarchical list. This process is repeated until all the different taxonomical classifications are present.

To replicate this XML format, several python scripts taking advantage of the "xml.etree.ElementTree" library were written to process and transform the CSV obtained from catalog of life and generate the correct syntax. Once the syntax was correctly replicated and the custom installation profile is ready, triggering the installation process again will prompt us with the listing of the installation profiles containing the custom created profile.

Once the system installs with the custom installation profile, the Figure 4.10 can be obtained in the lists menu:

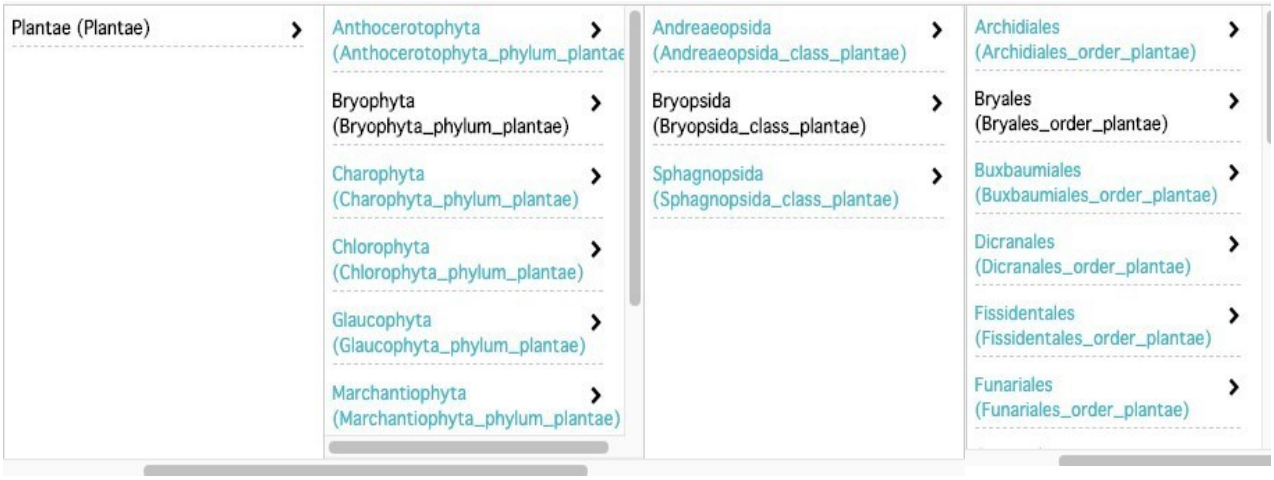

Figure 4.10- Representation of the taxonomic hierarchy using CollectiveAccess lists.

The Kingdom is in the first column. In this list only the Plantae kingdom is implemented, followed by every phylum and inside every phylum every class of that phylum and so on all the way down to specific epithet or infraspecific epithet if it exists.

4.7 Geography

When figuring out the best way to implement geographical functionalities, the same approach used for the taxonomy was considered, the creation of a list containing all possible countries, with their cities and divisions. After much consideration this approach was discarded, because that would result in a lot of values never used and the enormous amount of information necessary to create this type of hierarchical structure would be way to heavy on the installation profiles and on the database consuming a large amount of unnecessary space.
Fortunately, CollectiveAccess has a specific metadata datatype called “geonames”. This type of data takes advantage of an external API, more specifically from “geonames.org” where all we need to do is to start typing a localization, a city, a street or even a country and the external API returns a list of possibilities in the form of a dropdown we can choose from.
In the case of a correct match the coordinates are also stored and allow for their placing on a map for visualization.

An example is shown in Figure 4.11 where the word “beja” is typed in a metadata element with “geonames” as its datatype.

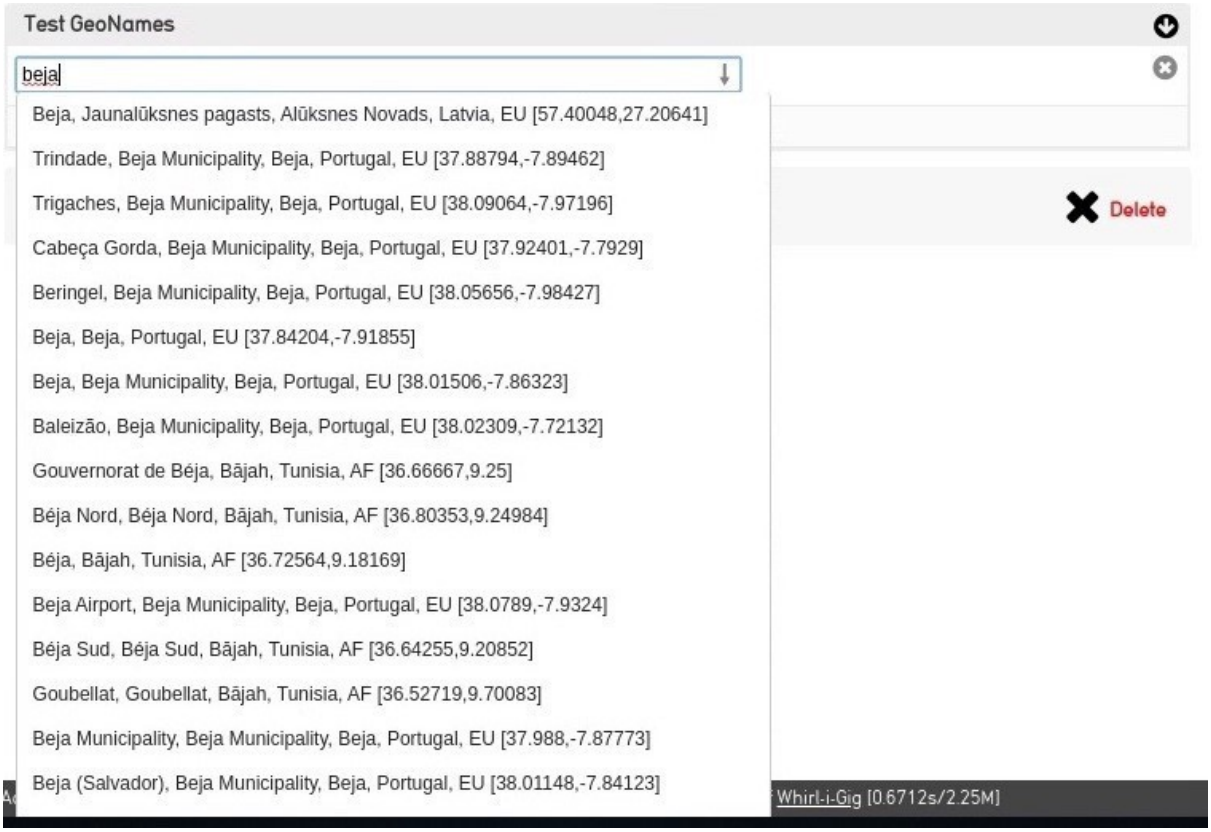

Figure 4.11- Example of a metadata element with type "geonames".

The more precise we are with the geographical information the shorter the lists of options get.

Using this method as the implementation also ensures that one of the problems stated in the requirements, with inferring information for geographical locations, is solved since all the information on the options will be stored as well.

4.8 Interfaces

Custom interfaces for different collections was also an important requirement. As the records' data is usually dependent on the curator, there's no universal insertion template that could be used for all data. An example of a custom interface is shown in Figure 4.12 for the LISC Herbarium Collection.

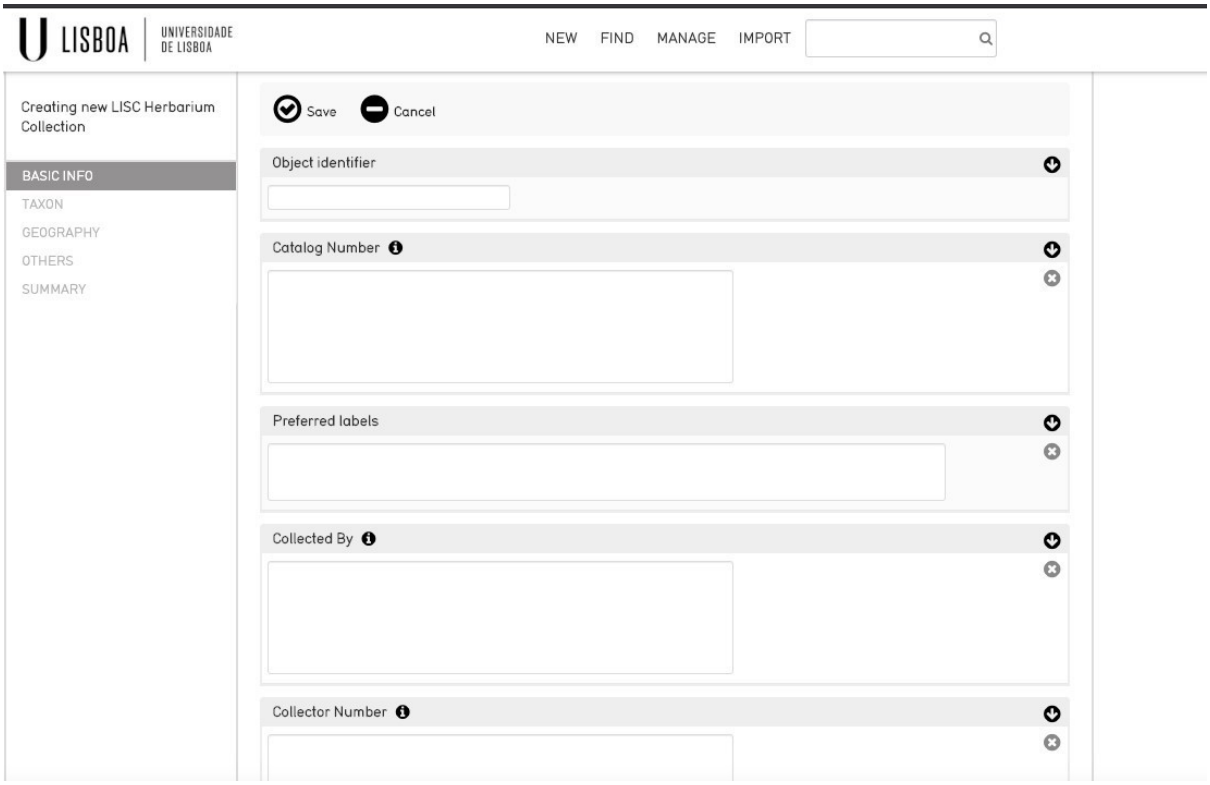

Figure 4 . 12 - Data Insertion interface for the LISC Herbarium Collection.

In this figure, some of the highlighted features mentioned in the state of the art (Section 2.2), about the cleanness and simpleness of the interface are present. Also, in the left part of this image there are different menus for this interface “Basic Info”, “Taxon”, “Geography”, “Others” and “Summary”. These menus allow for the organization of the metadata and instead of being a long web page with all the metadata to fill in a single page, the metadata is organized and placed inside the menus they relate to.

If the user were to click on the “Taxon” menu the following page would appear (Figure 4.13). Where the taxonomy related metadata, including this project’s implementation for the taxonomical features.

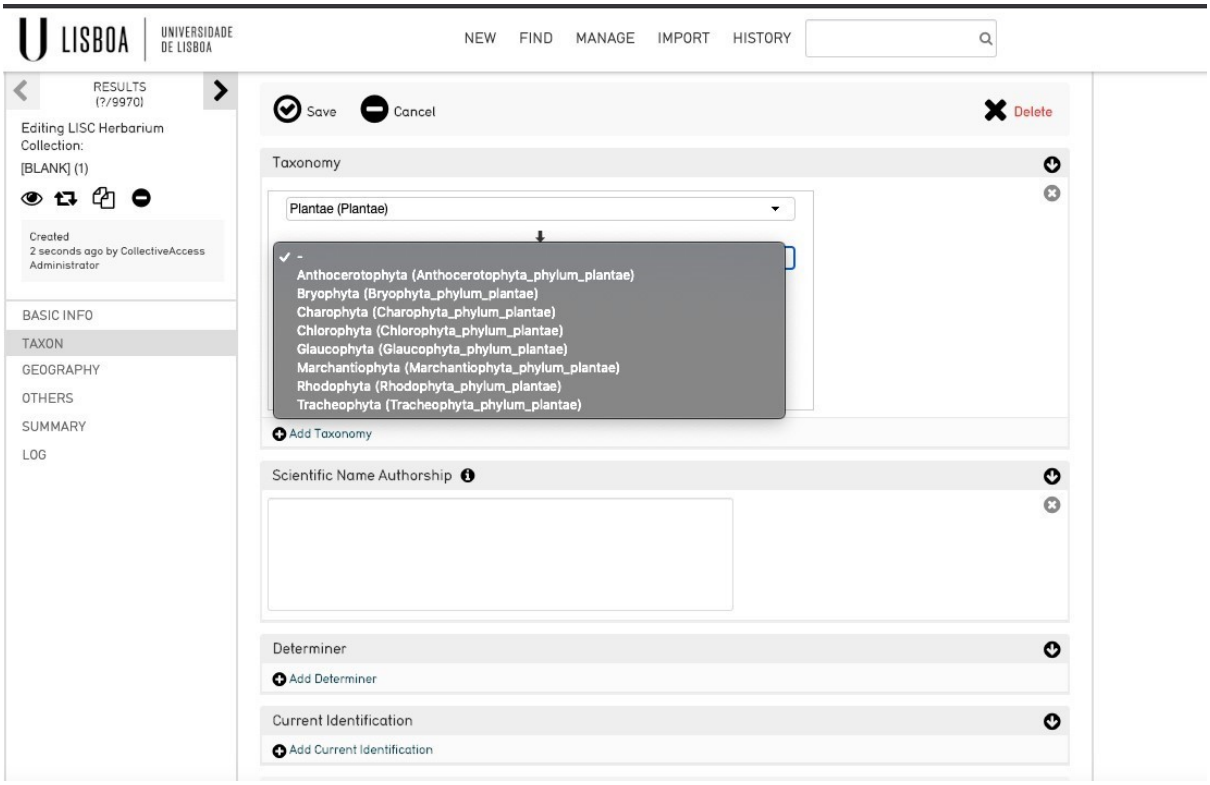

Figure 4.13- Data Insertion interface on the "TAXON" menu for the LISC Herbarium Collection.

5 Evaluation

User tests were performed to assess the performance and usability of the system. These tests were divided into two different categories with each one having a custom login with a custom set of permission and a different Google Form with instructions and operations to perform alongside some technical questions about the setups in terms of browsers, operating systems and screen resolution. These categories represented the two main roles for this web-portal, the public user role and the cataloguing and other data operations role performed by the curators. At the end of each form, suggestions about the usability and future improvements were asked, and a System Usability Scale (SUS) [12] questionnaire was included following the standard questions:

1- I think that I would like to use this system frequently.
2- I found the system unnecessarily complex.
3- I thought the system was easy to use.
4- I think that I would need the support of a technical person to be able to use this system.
5- I found the various functions in this system were well integrated.
6- I thought there was too much inconsistency in this system.
7- I would imagine that most people would learn to use this system very quickly.

8- I found the system very cumbersome to use.
9- I felt very confident using the system.
10- I needed to learn a lot of things before I could get going with this system.

SUS questions are evaluated on a 5-point Likert scale of strength of agreement. Its final score can range from 0 to 100, where higher scores indicate better usability [14].

Public User or Curators, authentication was not required in any way, making all users and their respective answers anonymous and the two Google Forms were provided simultaneous to them.

5.1 User tests

5.1.1 Public User tests

As part of the public usability test, a custom Google form was created to include instructions on the operations a public user with no affiliation with the Museums could have access to. These operations were exclusively search based operations and trying all the different types of searches available. The following results are based on a sample of 12 different people selected for their education in the Sciences of Life area, but with no experience in curating a collection.

Firstly, users were asked about their setup, so in case some performance issues were reported, correlations with specific software could be made. Out of the 12 test samples 3 people reported they were conducting the test on a fixed computer while 9 of them performed the test on a laptop. No person answered using a telephone or a tablet (Figure 5.1).

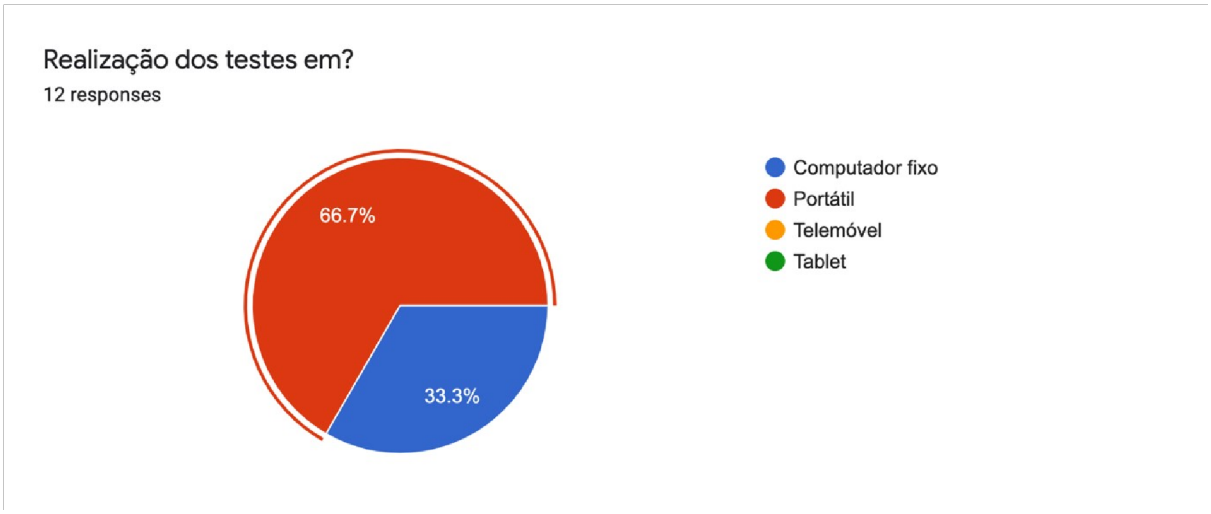

Figure 5.1- Results from the test environment regarding the physical hardware used for the public tests.

Users were also prompted for the number of screens they had and the browser the test was performed on (Figure 5.2), allowing for the verification of the website responsiveness throughout the most common browsers. With 9 answers, Google Chrome was the most used browser with most people having access to a single screen (9), which matches with number of people using a laptop.

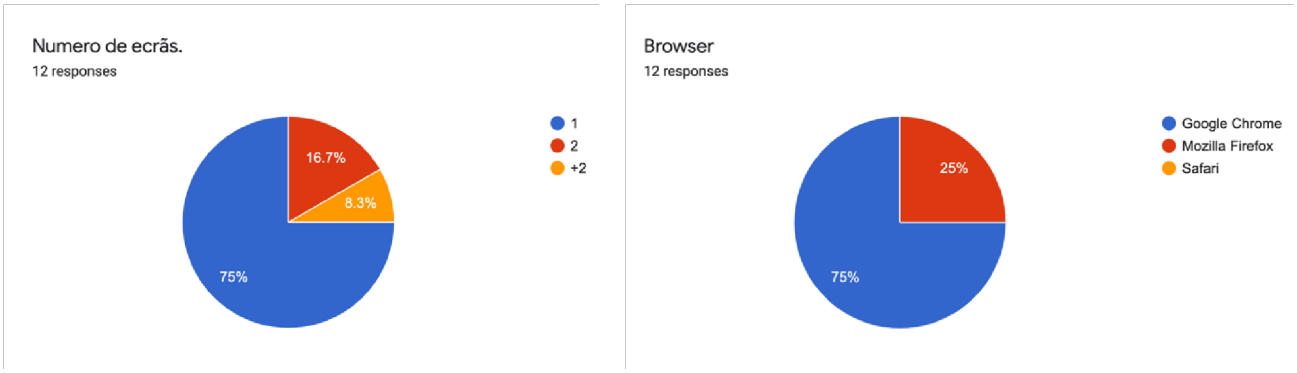

Figure 5.2- Results from public users when asked about their screen setups and web-browser used for the tests.

Amongst the search instructions given, users were asked to perform a simple search using a unique identifier of one of the records inside the system (Figure 5.3).

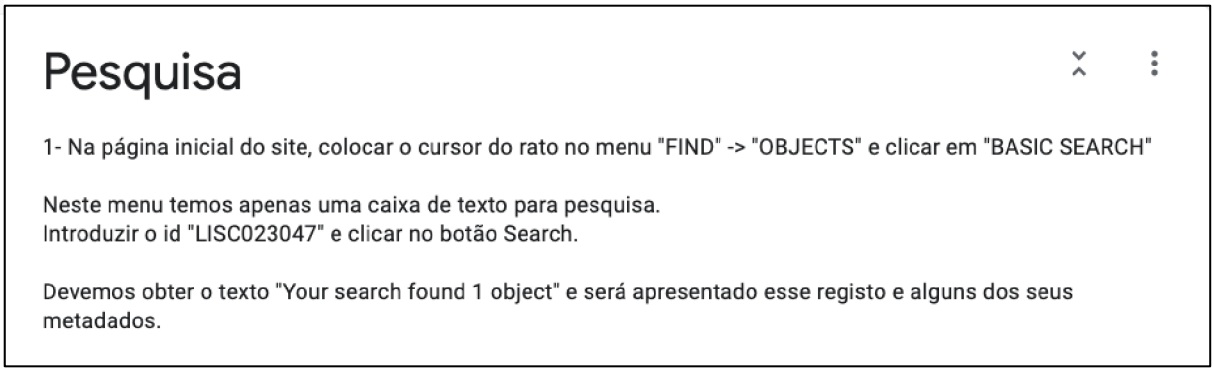

Figure 5.3- Google Form instructions for the simple search.

This is the most basic type of search as it is the most restrictive one and should only produce one record in the outcome. Users were asked to optionally send a screenshot via e-mail of this search to confirm the results, as having a file upload section would force authentication. Based on 7 e-mails all images contained exactly and only the single record their were asked to search. An image sent for this question is shown below (Figure 5.4).

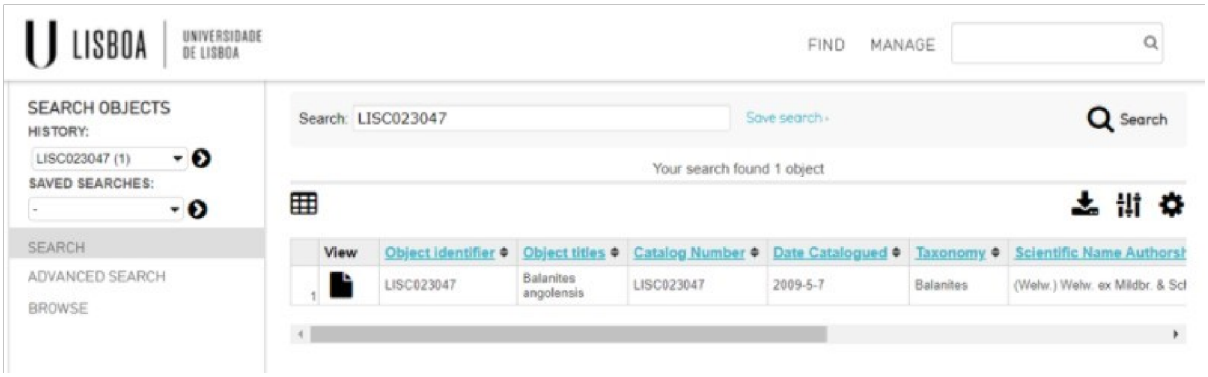

Figure 5.4- Example of an image received to confirm the search results.

Secondly, users were asked to try the advanced search interface, were multiple search fields were defined. They were asked to use the “Country” search field and search for all records in “Angola”. Here some small discrepancies to on the values obtained can be found, 9 of the users reported a search outcome of 9971 results, while 3 of them reported 9970 results (Figure 5.5).

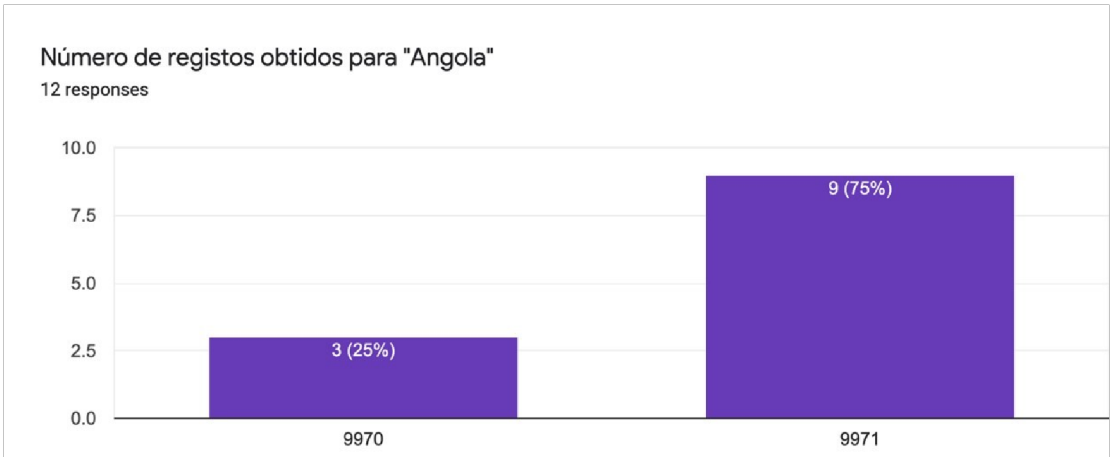

Figure 5.5- Number of records obtained from public users when performing a search with the keyword "Angola".

This can be explained when analyzing the curator's operations (see 5.1.2), that were happening simultaneous, where they were asked to insert a record that belonged to Angola and proceeded to fail the instructions on deleting the recently inserted record, adding up one more to the count.

On the topic of advanced search, users were also asked to narrow their searches to a more specific location, in this case a city in Angola called "Malanje". As it is a more restrictive search the results should be lower, because all "Malanje" records should appear in the "Angola" search but not all "Angola" records belong to "Malanje". The outcome of this search results was unanimous with all 12 people answering 203 as the number of obtained records (Figure 5.6).

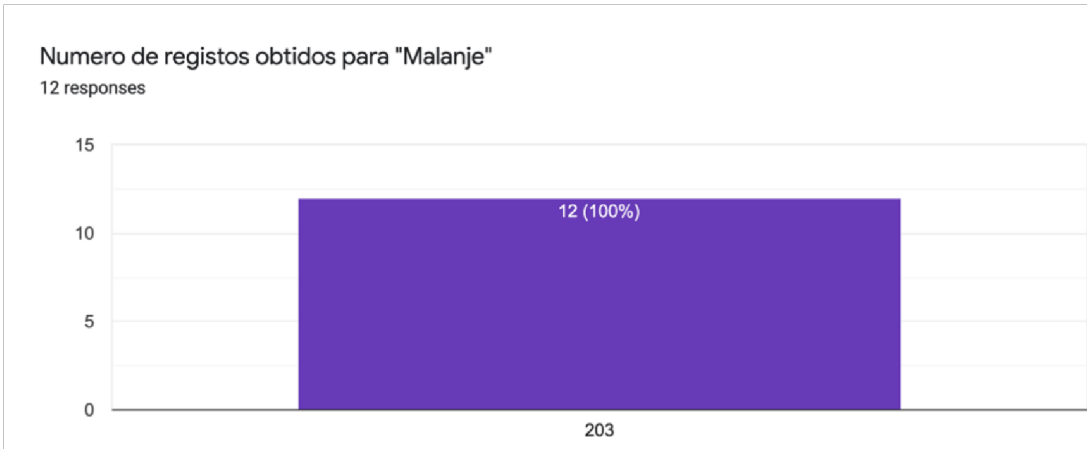

Figure 5.6- Number of records obtained from public users when performing a search with the keyword "Malanje".

Lastly, users were asked to use the "Browse" search functionalities were the taxonomical searches were implemented. In this search, a full list of all the taxonomical classifications of the records in the system was presented in alphabetical order, and users were asked to search for all records of genus

"Balanites". The responses were unanimous with all 12 participants answering 37 search results (Figure 5.7).

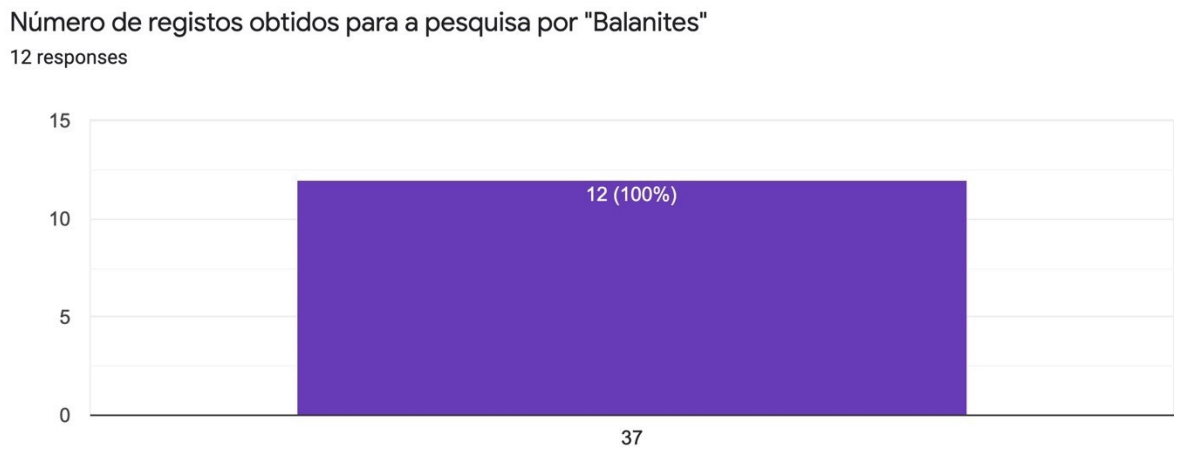

Figure 5.7- Number of records obtained from public users when performing a browse search with the keyword "Balanites".

After these search operations the SUS questionnaire was answered, and the following responses were obtained (Figure 5.8).

Users	Question 1	Question 2	Question 3	Question 4	Question 5	Question 6	Question 7	Question 8	Question 9	Question 10
User 1	5	1	5	1	4	1	4	1	5	2
User 2	5	1	5	1	5	5	4	1	5	1
User 3	5	1	5	1	4	1	5	1	4	1
User 4	5	1	5	1	5	1	5	5	5	1
User 5	4	2	4	3	4	2	4	2	4	4
User 6	4	1	4	1	4	1	4	1	5	2
User 7	5	1	4	2	5	1	4	1	4	1
User 8	3	3	3	2	4	2	3	2	3	3
User 9	4	2	4	2	4	2	4	2	4	2
User 10	4	2	3	3	4	2	4	2	4	2
User 11	5	2	4	2	4	2	4	2	3	3
User 12	4	2	4	3	4	2	4	2	4	2

Figure 5.8- Individual answers to the SUS questionnaire from the public tests.

The overall SUS scores were calculated following the “ODD question value -1” + ”5- pair question value” to convert all values to a [0-40] range and then multiplied by 2.5 to give it a [0-100] range (Figure 5.9). The obtained average score for the public users was 80 and the standard deviation was 11.9.

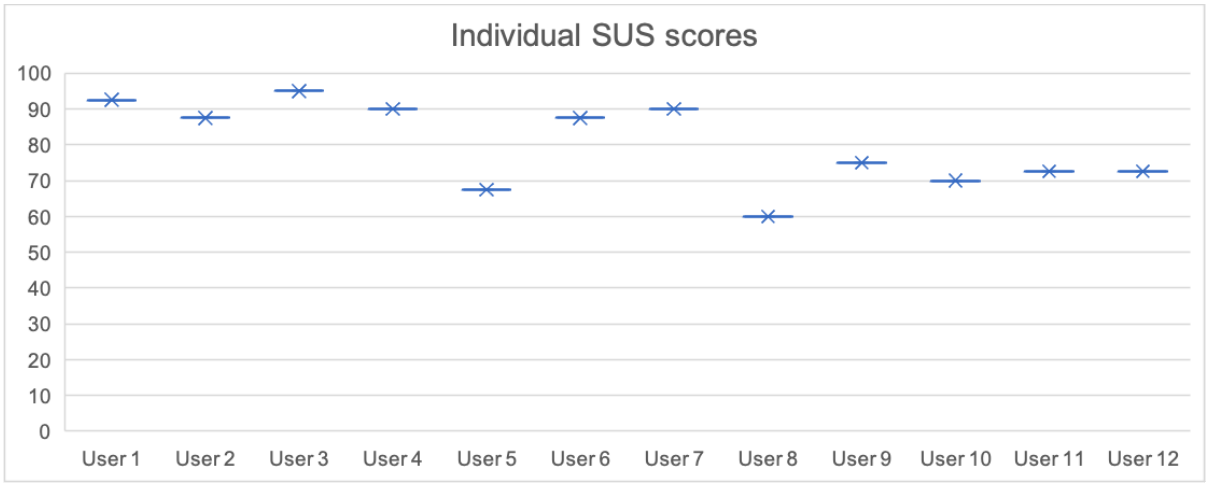

Figure 5.9- Individual SUS scores from the public tests converted to a scale from 0-100.

At the end of the questionnaire, users were asked to provide any suggestions or bugs they found while using the interface. No responses regarding the bugs were obtained, even if the tests were conducts on only one screen, and likely that screen had to be divided into two to have the Google form side-by-side with the

web-portal. One suggestion was obtained where the user suggested that when using the browse feature for searching records via their taxonomical classification, the listing that appears should be split into the different taxonomical groups, for instance, have a list for genus browsing only, for family only, etc.

5.1.2 Curators test

Unlike regular users, curators are allowed to perform data operations. For these tests besides the search operations which were the same as the public users, curators were also asked to insert a specific record with the information being provided in their Google form, using the basic search to find that record, and at the end delete the record. The following analysis is based on a sample of 10 curators who answered the questionnaire, carefully selected for either being a curator at the MUHNAC or in other museum/university or for having some experience with herbarium database management.

The test began with the same setup questions about the personal setups to see if there were performance or web formatting issues that could be attributed to a software choice.

Out of the 10 curators, 6 reported they conducted the operations on a fixed computer while 4 curators did it on a laptop (Figure 5.10).

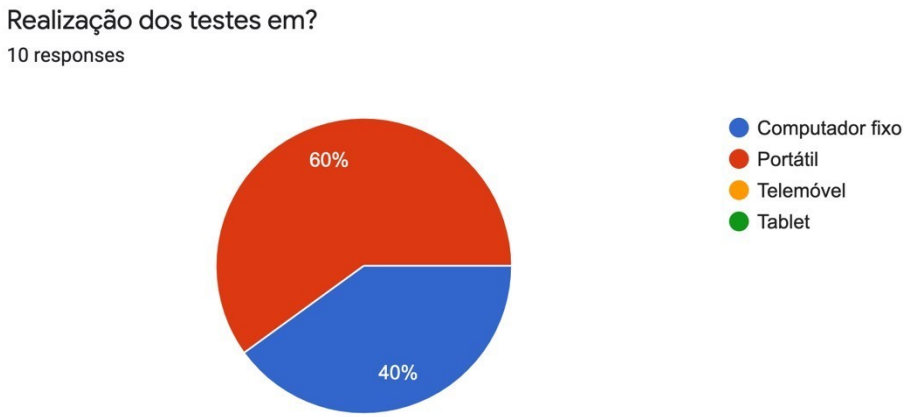

Figure 5.10- Results from the test environment regarding the physical hardware used for the curator tests.

The number of screens and browser (Figure 5.11) used to conduct the test follow the same pattern as the public test, with the majority using Google Chrome (7) as their browser and only having access to a single screen (7). A surprise here comes from the fact one curator reported to have conducted the test using Brave as their browser. This was the only browser from the list that was not tested in any way while implementing the system.

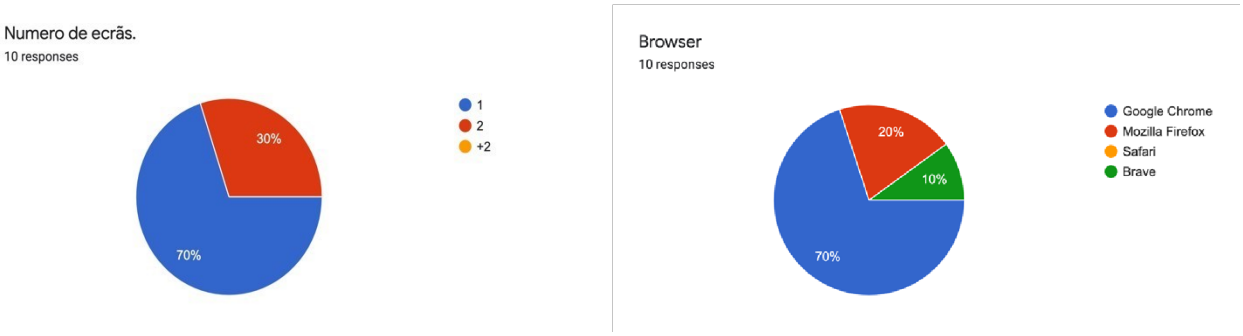

Figure 5.11- Results from the curators when asked about their screen setups and web-browser used for the tests.

Following the setup questions, the insertion of a new record belonging to the LISC Herbarium Collection was the next operation. Curators were told, via the form, all the information they needed to use to fill the insertion interface and the instructions on how to navigate the screens (Figure 5.12, Figure 5.13 and Figure 5.14).

Inserção

1- Na barra de navegação vamos pôr o cursor em cima de "NEW" e seguir os caminhos: "Object" -> "Natural Sciences" -> "Botany" -> e clicar em "LISC Herbarium Collection".

Devemos ter uma página com caixas de texto para preencher e no lado esquerdo da página deverá dizer "Creating new LISC Herbarium Collection". Se não for o caso, repetir o passo 1.

De realçar que, também no lado esquerdo da página, podemos ver 5 menus: "BASIC INFO", "TAXON", GEOGRAPHY", "OTHERS" e "SUMMARY" que servem para distinguir e agrupar diferentes atributos de um mesmo objecto. Neste momento encontramo-nos no menu "BASIC INFO".

Vamos preencher as caixas de texto do menu inicial "BASIC INFO" com os seguintes valores:

Object identifier : LISC024173
Catalog Number : LISC024173
Preferred labels : Melia azedarach
Collected By : C.Henriques
Collector Number : 118
Start Time : 1963-02-13
Access : accessible to public(opção do dropdown)
Status : new(opção do dropdown)

Após estes dados estarem preenchidos vamos clicar no botão de "SAVE" no final ou no início da página.

Figure 5.12- Google Form with the instruction for the record insertion (1/3).

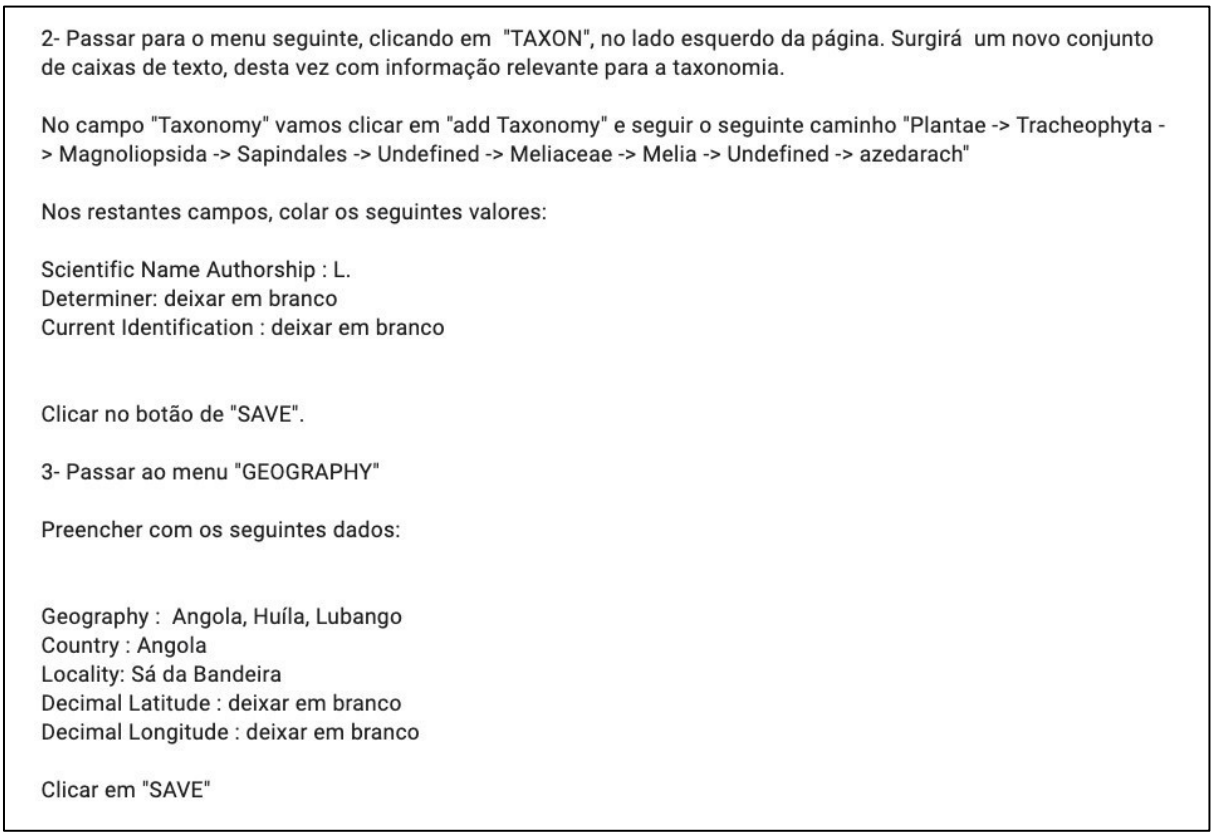

Figure 5.13- Google Form with the instruction for the record insertion (2/3).

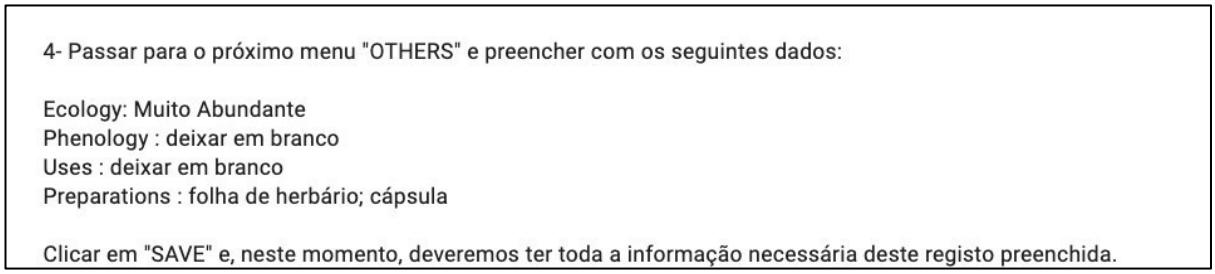

Figure 5.14- Google Form with the instruction for the record insertion (3/3).

Afterwards, users were asked to provide feedback on the difficulty of using the software (Figure 5.15), with 5 people answering they found the system and the interfaces very easy to use, 3 people reported it as just being easy, while 2 people found it hard, and leave suggestions for improvements.

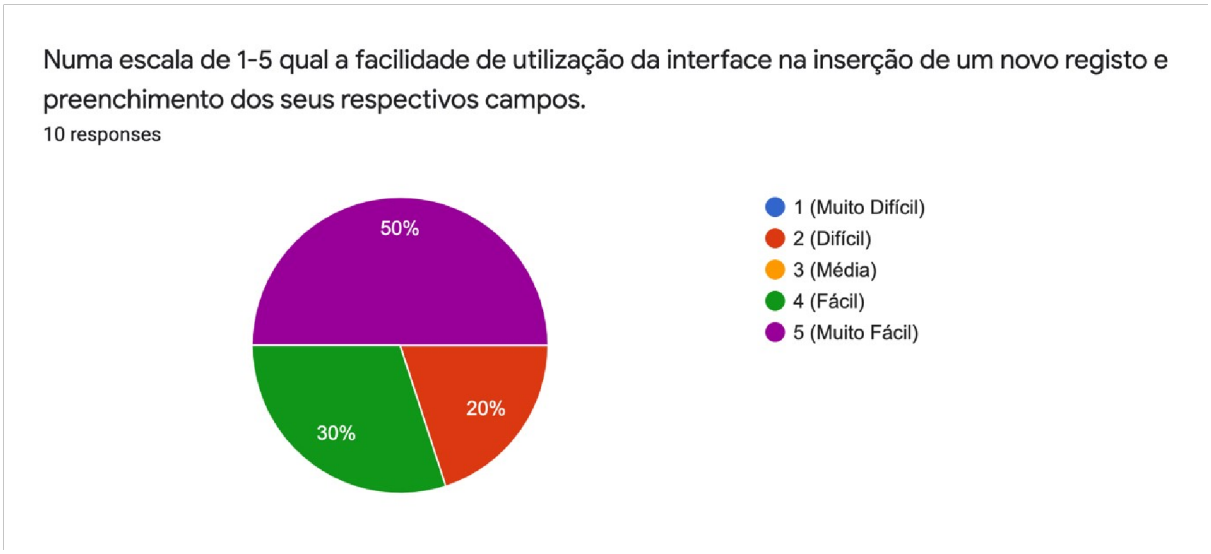

Figure 5.15- Difficulty in using the software interfaces and filling the forms.

Four suggestions were made, with two of them focused on the increase of controlled vocabulary metadata elements for fields like "Country", "Collected by", "Preparations" to minimize errors and to guarantee consistency across all records. The remaining two suggestion were concerned about the filling of the taxonomy metadata, which in the current implementation users have to follow the hierarchy provided and navigate the hierarchical structure all the way to the specific epithet, instead of the field being a plain text field were the species name is typed and behind the scenes all the above classifications are inferred.

The search operations performed were the same as the public test ones, with no deviations from what was expected as the information for the newly inserted record did not coincide with any of the search operations, apart from some users reporting that when they searched for the record they had just introduced they would see multiple of them (Figure 5.16).

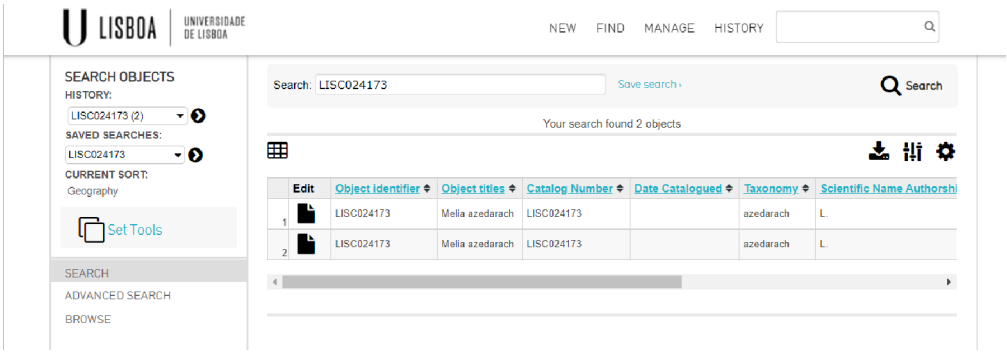

Figure 5.16- Image received via e-mail showcasing duplicate record with the same identifiers.

Several reasons can cause this to happen. Either 2 curators did the test simultaneous and when this search operation was performed, multiple records with the same information were still present in the system, since the data used in the insertion step was the same for all curators. A different reason, and perhaps the most likely to have happened, is that the step in which users were asked to delete the record they introduced was not completed successfully, leading to the record staying inside the system and causing it to show multiple times during searches by identifiers. This was an expected problem from the start

since it was decided that for the usability tests, the duplication of unique identifiers would be allowed. If the duplication of unique identifiers was not allowed for the tests in the case of a curator failed the instructions on how to delete the record, the following curators to answer the form would have not been able to follow the insertion instructions as they would get an error message. As soon as the tests were over, this configuration was changed and if anyone tried to insert a record using a unique identifier already present in the system, a warning would be shown, and that record would not be inserted.

As with public users, curators were also prompt with the SUS questionnaire with their answers being the follow.

Users	Question 1	Question 2	Question 3	Question 4	Question 5	Question 6	Question 7	Question 8	Question 9	Question 10
User 1	2	1	4	2	4	2	4	1	4	1
User 2	5	1	5	1	5	1	5	1	5	1
User 3	5	2	5	1	4	1	5	1	4	2
User 4	3	3	4	4	3	3	4	2	3	1
User 5	5	1	4	2	5	1	3	1	3	1
User 6	5	1	5	1	5	1	5	1	5	1
User 7	4	2	4	2	4	3	4	2	3	5
User 8	4	1	5	1	4	1	5	1	4	1
User 9	4	2	4	3	4	3	4	2	3	3
User 10	2	3	3	3	2	4	3	3	2	3

Figure 5.17- Individual answers to the SUS questionnaire from the curator tests.

Using the same formula to analyze the SUS results and convert them into the [0-100] range the individual scores were obtained (Figure 5.14). The average score for the curator's questionnaire was 77,25 while the standard deviation was 19,84, in part because of the answers provided by User 10.

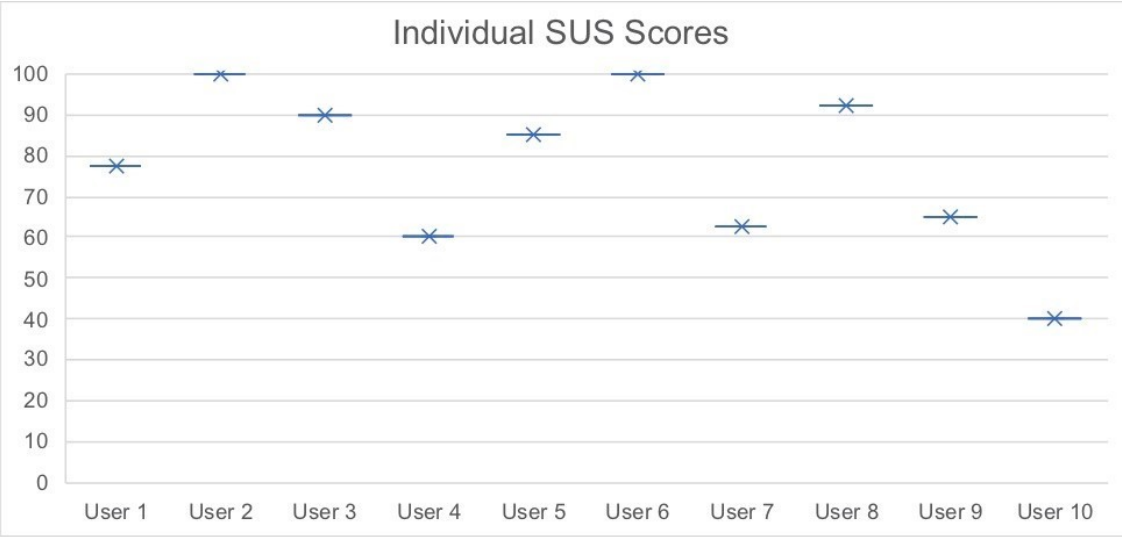

Figure 5.18- Individual SUS scores from the curator tests converted to a scale from 0-100.

At the end of the questionnaire curators were asked to provide an overall feedback and include suggestions about the operations performed and to include suggestions for the future continuation of this project.

The following suggestions were obtained, with the focus being on:

- Increase use of controlled vocabulary fields and to try to use it as much as possible.

- Taxonomical related operations being divided into their own isolated searches

- Allowing for the introduction of only the name of the species and leaving the software to figure out the rest.

- Possibility to observe when available the images associated with the records.

5.1.3 Overall Evaluation

Considering both suggestions and results from the public tests and the curator's tests, there is still a lot of improvements that need to be made in order for this system to became fully operational.

No interface bugs were found or reported, which was pleasing since the original software portal was not responsive causing some elements like buttons and information to stay off-screen if the page was shrank to a point, and changes to the styling files (CSS) had to be made to include media queries to increase the responsiveness.

The taxonomical functionalities, either when inserting a record or searching, needs to be revamped and the feedback on the correct way it should behave will most definitely be taken in consideration for future developments.

The increase of controlled vocabulary fields is something that was very suggested, and even though the version of the software used for the tests did not include many types of fields, it is definitely something that was discussed internally as an improvement, and preparations had already begun to add use this type of fields as much as possible.

The visual representation of the records was also not included in the testing version but, like controlled vocabulary list, is already under development with a small percentage of the records having their images associated with them, and the possibility of searching for those specifically is already implemented.

The SUS questionnaire results from the public tests produced an average score of 80 and a standard deviation of 11.9, while the curator's tests had an average of 77,25 and a standard deviation of 19,84. These results rank our system with the A grade in terms of the system usability for the public testing and a B score when it comes to the curators, following the classification presented in Figure 5.19.

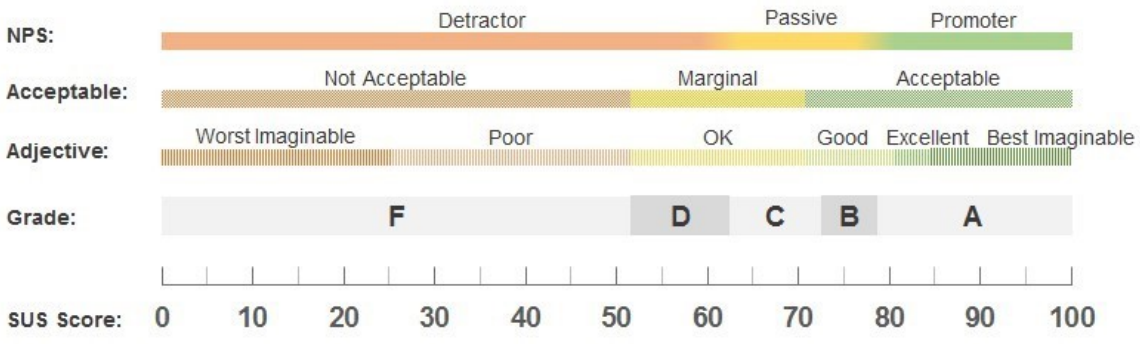

Figure 5.19- A comparison of the adjective ratings, acceptability scores, and school grading scales, in relation to the average SUS score.

6 Conclusions and Discussion

This project designed and implemented a first approach of a system to store, manage and search scientific collections for MUNHAC with a specific focus in the biological domain. The challenges of integrating data belonging to multiple domains of the Sciences of Life and Natural History areas in a single platform, while at the same time meeting all the requirements and complexity that data brings is a major priority for scientific collection institution holders. To attempt at this, this system was designed on top of an open-source software named CollectiveAccess (CA) which provides a web front-end for both the general public and MUNHAC curators, amongst some other core functionalities. Using the built-in features of CA complemented with this project's implementations and extensibility of those functionalities, this work managed to provide a platform which provides support for manual data insertion through the use of customizable interfaces with custom metadata elements to fit the needs of every collection. It also provides a way to bulk import data through the use of the mappings files already build and the association of images, videos and files to the data, as well as batch editing of records to perform data maintenance whether it is to update a value on a set of records or to delete multiple records chosen on a common property.
Search operations can be individualized for every collection with the possibility of performing the most basic search operation through a unique identifier, to the utilization of the advanced search features where search forms can be customized for the curators, collection managers, conservators, containing multiple searchable metadata elements and allowing for joint searches based on different data properties, and browse searches which provides a way for more exploratory searches.
Data exportation can be achieved simply by performing any type of search and pressing the download icon to export the search results to the user computers in any of the main file extensions (e.g., CSV, XSLX, PDF).
Taxonomical and geographical information are supported via the use of a controlled vocabulary list and using an external API respectively, with the possibility of geolocalization of records on a map representation, and based on the test user's feedback, controlled vocabulary lists will be used as much as possible in the future.
Custom permissions allow for safety inside the system ensuring each collection can only be managed for the curators or other authorized people responsible by them.
Loan/borrows support and object lots support began to be tested but are still pending on official documentation for further testing.
To keep these features and the data secured, two measures of backups are in place. One of them utilizes the installation profiles, where all structural elements of the system were replicated and can be automatically generated if needed, and regular databases backups.

The functionalities implemented in this project provided a first approach to implement a collection management portal for the collections of the MUHNAC regardless of their type. However, special modules were implemented to support biological collections. Future works can seek to improve the system by:

- Increasing the number of collections inside the system, building custom metadata elements and interfaces as needed.
- Improving the taxonomical functionalities accordingly to the users' feedback.

- Increasing the number of metadata elements associated with a controlled vocabulary list to minimize possible errors and ensure consistency.
- Improving the Loan support features, and other management processes, by mirroring the MUHNAC official loan documentation.
- Contributing to the process of digitalization of records and upload them to CollectiveAccess.
- Extend user tests to gather more feedback.

Analyzing the overall performance of CA, considering the usability tests, the internal tests and the opinions of the MUHNAC curators, CA has proven to be a powerful software capable of answering or providing a way to answer nearly every specificity that a particular record might have while at the same time having a simple and somewhat intuitive interface layout. All these inputs together provided valuable information regarding the implementation of the functionalities in this work, suggestions for improving the same functionalities as is the case of the taxonomical features, and ideas for further features.

CA is also platform that is used widely internationally with some academic institutions like the European University Viadrina in Frankfurt (Oder), Germany; the School of Visual Arts in New York using it as their platform. Museums like the FeliXart Museum in Drogenbos, Belgium; Musée Chappuis-Fähndrich in Develier, Switzerland amongst other institutions, museums, libraries, historical societies also employ CA as their back-end engine. Besides theses international institutions even in Portugal projects using CA can be found, as is the case of the "Plataforma dos Açores Digital" and the Museum of the University of Aveiro (MUSA) proving yet another proof of the capabilities and flexibility of this software throughout different collection types.

To conclude, this work is a contribution to make the data about the MUNHAC collections FAIR (findable, accessible, interoperable and reusable) by providing a single system to store all collections data, that follows international metadata standards (e.g., Darwin Core, Catalogue of Life) and is able to export in the formats of other relevant efforts (e.g., GBIF). It is thus expected that this work will not only greatly facilitate the integrated management of collections at MUNHAC, but also support valuable uses of this data for further scientific effords.

7 Bibliography

[1] Fillinger, S.; de la Garza, L.; Peltzer, A; *et al.* Challenges of big data integration in the life sciences; *Anal Bioanal Chem* 411**;** 6791–6800 (2019); DOI: 10.1007/s00216-019-02074-9

[2] National Science and Technology Council, Committee on Science, Interagency Working Group on Scientific Collections. Scientific Collections: Mission-Critical Infrastructure of Federal Science Agencies; Office of Science and Technology Policy; Washington, DC. 2009

[3] Holetschek, J.; Baumann, G.; Koch, G.; Berendsohn, W.G.; 2016; Natural History in Europeana - Accessing Scientific Collection Objects via LOD. In: E. Garoufallou et al. (Eds.): MTSR 2016; CCIS 672, pp. 223–234, 2016. DOI: 10.1007/978-3-319-49157-8_20

[4] Scott, E.; Baker, E.; Woodburn, M.; Vincent, S.; Hardy, H.; Smith, V.; 2009; The Natural History Museum Data Portal; Database; DOI: 10.1093/database/baz038

[5] Bakker, F.T.; Antonelli, A.; Clarke, J.A.; Cook, J.A.; Edwards, S.V.; Ericson, P.G.P.; Faurby, S.; Ferrand, N.; Gelang, M.; Gillespie, R.G.; Irestedt, M.; Lundin, K.; Larsson, E.; Matos-Maraví, P.; Müller, J.; von Proschwitz, T.; Roderick, G.K.; Schliep, A.; Wahlberg, N.; Wiedenhoeft, J.: Källersjö, M.; 2020. The Global Museum: natural history collections and the future of evolutionary science and public education; PeerJ; DOI: 10.7717/peerj.8225

[6] OECD (1999). Final Report of the OECD Megascience Forum Working Group on Biological Informatics. 74 pp

[7] Horrocks, I.; 2018; Ontologies and the semantic web; Communications of the ACM; 51(12):58-67; DOI: 10.1145/1409360.1409377

[8] Hoehndorf, R.; Schofield, P.N.; Gkoutos, G.V.; The role of ontologies in biological and biomedical research: a functional perspective; 2015; Briefings in Bioinformatics; 16(6):1069-1080 ; DOI: 10.1093/bib/bbv011

[9] Access to Biological Collection Data task group. 2007. Access to Biological Collection Data (ABCD), Version 2.06. Biodiversity Information Standards (TDWG)

[10] Darwin Core (DwC). Retrieved from http://rs.tdwg.org/dwc/terms.htm

[11] NWO Research Fields. Retrieved from https://www.nwo.nl/en/nwo-research-fields

[12] Brooke, J.; 1996; SUS: a “quick and dirty'usability; Usability evaluation in industry; p.189.

[13] Catalog of Life. Retrieved from http://www.catalogueoflife.org/DCA_Export/

[14] Martins, A. I.; Rosa, A.I.; Queirós, A.; Silva, A.; Rocha, N.P.; European Portuguese Validade of System Usability Scale (SUS); 2015 ; Procedia Computer Science ; 67:293- 300 ; DOI:10.1016/j.procs.2015.09.273